Master in Management + Analytics

Predictive analytics in legal billing: an applied machine learning approach to forecasting bill rejections

Autoría: Quevedo, Carlos María
Año: 2025

UNIVERSIDAD TORCUATO DI TELLA

MASTER IN MANAGEMENT + ANALYTICS

PREDICTIVE ANALYTICS IN LEGAL BILLING: AN APPLIED MACHINE LEARNING APPROACH TO FORECASTING BILL REJECTIONS

THESIS

Carlos María Quevedo

May, 2025

Advisor: Viviana Siless

Abstract

In the legal industry, accurate billing is not only a matter of financial importance but also critical to maintaining strong client relationships. However, bill rejections or discounts requested by clients can significantly impact the revenue streams of law firms. This thesis presents a practical application of machine learning -the XGBoost algorithm- to forecast the likelihood of bill rejections based on historical billing data. The research explores various factors that contribute to bill rejections, including project rates, billing office attributes, employee roles, and the narratives associated with the work descriptions.

Through the development and deployment of a predictive tool, this thesis provides a data-driven approach to identifying high-risk bills before they are sent to clients. The findings suggest that while narratives are important, other factors such as project area and billing office play a more significant role in determining whether a bill will be accepted or rejected. This work also delves into the preprocessing techniques, feature engineering, and hyperparameter optimization processes that are crucial to the model's success. The implications of these findings are discussed in the context of improving legal billing practices and reducing financial risks for law firms.

Index

Table Index

Figure Index

1. Introduction

1.1 Overview of Legal Billing and its Challenges

Legal billing is a fundamental process that underpins the financial health of law firms. It is a detailed and often complex task that involves accurately recording the time and resources spent on behalf of clients and then converting this information into invoices. However, the billing process in law firms is not merely about compiling and sending invoices. It encompasses a range of activities, including time tracking, cost allocation, invoice preparation, and client communication. Each of these activities must be executed with precision to ensure that the final invoice reflects the true value of the services provided.

The legal billing process typically begins when a law firm takes on a new client. Throughout the engagement, billable time and expenses are meticulously logged. At the end of each billing cycle, usually monthly or at the conclusion of a case, these entries are compiled into draft bills. Attorneys then review these drafts, making necessary adjustments to ensure accuracy and completeness. Once finalized, the bills are sent to clients for payment.

Despite its structured approach, the legal billing process is fraught with challenges. According to the 2020 Legal Trends Report (Clio, 2020), lawyers record only 2.5 billable hours per day on average (31.25% utilization rate), with the rest of their day consumed by non-billable tasks, many of which are related to billing. These tasks include preparing invoices, processing payments, and following up on unpaid bills. The time and effort required to manage these tasks can detract from a lawyer's ability to focus on billable work, ultimately affecting the firm's profitability.

Moreover, the billing process is prone to various bottlenecks. Delays can occur at multiple stages, such as when attorneys take too long to approve bills or when there are discrepancies in the billing descriptions that need to be resolved. Delays can mean late payments, cash flow issues, and even client dissatisfaction.

One of the most significant challenges in legal billing is managing disputes and bill rejections. Clients may reject bills for several reasons, including discrepancies between the billed

hours and the work performed, ambiguous or insufficient descriptions of tasks, or billing rates that do not align with their expectations. These rejections can lead to lengthy disputes, which not only delay payments but also strain client relationships.

1.2 Importance of Accurate and Transparent Billing

Accurate and transparent billing is crucial for maintaining trust between law firms and their clients. Clients need to feel confident that they are being billed fairly for the services provided. This confidence is built through clear, detailed billing descriptions and consistent communication throughout the billing process.

Legal billing descriptions should provide sufficient context to justify the charges. For instance, instead of a vague entry like "Legal research," a more detailed description would be "Conducted legal research to support the motion for summary judgement." Such clarity helps clients understand the value they are receiving, reducing the likelihood of disputes.

In addition to clear descriptions, law firms must also have standardized billing policies in place. These policies ensure consistency across all invoices and provide a framework for handling any billing issues that arise. For example, a well-defined billing policy would specify when invoices should be sent, how to handle billing disputes, and the procedures for following up on late payments.

However, even with the best billing practices in place, law firms often face challenges in ensuring that their bills are accepted by clients without dispute. This is where predictive analytics can play an important role.

1.3 Analytics in Legal Billing

Predictive analytics involves the use of statistical models and algorithms to analyze historical data and make predictions about future outcomes. In the context of legal billing, predictive analytics can be used to forecast the likelihood of bill rejections based on patterns observed in past billing data. This proactive approach allows law firms to identify potential issues before they arise, enabling them to take corrective action and minimize the risk of disputes.

The application of predictive analytics in legal billing represents a significant shift from traditional methods, which are often reactive in nature. Traditionally, law firms would only address billing issues after they have been flagged by clients, after which the damage -both financial and relational- may already have been done. By contrast, predictive analytics enables firms to assess the risk of bill rejection at the time of invoice preparation, allowing them to make adjustments that increase the likelihood of client acceptance.

For example, if the predictive model identifies that invoices with certain types of narratives are more likely to be rejected, the firm can amend these narratives to provide more detail or clarity. Similarly, if the model suggests that certain billing rates are frequently disputed, the firm might reconsider its pricing strategy for those services.

1.4 Development of the Predictive Model

The core of this study revolves around the development of a predictive model designed to forecast bill rejections. The model leverages the XGBoost algorithm, a powerful machine learning technique known for its accuracy and efficiency in handling large datasets. XGBoost is particularly well-suited for this application because it can capture complex interactions between variables, making it ideal for analyzing the multifaceted data involved in legal billing (Chen & Guestrin, 2016).

The model development process began by collecting and preprocessing historical billing data from a leading global law firm with thousands of attorneys and hundreds of offices worldwide. This firm provides legal services to multinational corporations, financial institutions, and government agencies, covering practice areas such as corporate transactions, mergers and acquisitions, dispute resolution, tax, intellectual property, labor and employment, and regulatory compliance. The dataset includes details on the type of legal service provided, the amount billed, the specific narratives used to describe the work performed, and whether the client accepted or rejected each bill. To enhance predictive performance, the process involved engineering additional features through polynomial transformations of numerical variables and clustering of

narrative descriptions to identify meaningful patterns.

A key aspect of the data preprocessing involved the vectorization and tokenization of the narrative descriptions. Legal narratives are often unstructured text data, which poses a challenge for predictive modeling (Manning et al., 2008). This was addressed by converting the narratives into categorical features using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. This approach allowed the model to capture the importance of specific words and phrases within the narratives, providing insights into how different descriptions impact the likelihood of bill rejection.

1.5 Feature Importance and Model Evaluation

One of the critical outputs of the predictive model is identifying the most important features that influence bill rejections. Feature importance analysis reveals which variables have the greatest impact on the model's predictions, providing valuable insights into the factors that drive client acceptance or rejection of bills.

The study involved analyzing narrative descriptions to find out their relative importance and if they are significant predictors along with other factors, such as billing rates and type of legal services provided. The findings presented in section 4.3 underscore the complexity of the billing process and highlight the need for a multifaceted approach to managing billing practices.

The evaluation of the model involved various performance metrics, including accuracy, precision, recall, and the area under the ROC curve (AUC). These metrics offer a comprehensive view of the model's effectiveness, helping assess if it can reliably predict bill rejections in real-world scenarios (Powers, 2011).

A user-friendly tool that law firms can use to input new billing data and receive predictions on the likelihood of bill rejections integrates the final model into a practical application. This tool is designed to be easy to use, requiring no technical expertise on the part of the user. By providing actionable insights in a straightforward format, the tool empowers law firms to proactively manage their billing processes and reduce the incidence of disputes.

1.6 The Evolution of Legal Billing Practices

Over the years, legal billing practices have evolved significantly. Traditionally, law firms relied heavily on billable hours as the primary method of charging clients for legal services (Thomson Reuters, 2023). This method, while straightforward, often led to criticism regarding inefficiency and lack of transparency. Clients, particularly those from large corporations, began demanding more clarity and clearer reasoning for billed hours, leading to the development of alternative billing arrangements.

The emergence of alternative billing models, such as flat fees, contingency fees, and subscription-based services, has been driven by the need to align legal costs more closely with the value provided to clients. These models offer greater predictability and transparency, which are increasingly important in today's competitive legal market. However, the complexity of legal work means that even these alternative models require careful management to ensure that charges to clients remain accurate and reasonable

In this context, the integration of technology into legal billing practices has become increasingly important. Legal practice management software, time tracking tools, and billing automation systems are now widely used to streamline the billing process and reduce the administrative burden on law firms. These tools not only improve efficiency but also enhance the accuracy of billing, helping to prevent disputes, and ensuring that clients are billed fairly for the services they receive.

1.7 The Role of Machine Learning in Modern Legal Practices

Machine learning is transforming a wide range of industries, and the legal sector is no exception (Susskind, 2017). By automating complex tasks and providing data-driven insights, machine learning technologies are helping law firms to operate more efficiently and make better decisions. In the context of legal billing, machine learning can be used to analyze large volumes of data, identify patterns, and make predictions that would be difficult or impossible for humans to achieve on their own.

In the legal billing domain, a significant portion of the available information comes in the form of unstructured data, such as the narrative descriptions that accompany invoices. This type of data is often highly variable and context-dependent, making it difficult for traditional analytical methods to process effectively.

In contrast, artificial intelligence and machine learning algorithms are well-equipped to handle unstructured data, as they can detect complex patterns and relationships that are not immediately apparent, facilitating the uncovering of valuable insights.

The development of the predictive model in this thesis applies machine learning techniques to analyze both structured data (such as billing rates and service types) and unstructured data (work descriptions). Combining these data types enables a more comprehensive analysis and captures the full complexity of the factors that influence bill rejections.

Moreover, machine learning models can continuously learn and improve over time as they are exposed to new data. This means that the predictive model developed has the potential to become even more accurate as it is used in practice, providing law firms with increasingly reliable insights into their billing processes.

1.8 Structure

This thesis is organized into several chapters, each addressing a different aspect of the research and its findings. Following this introduction, Chapter 2 provides a short review of the literature on legal billing practices, predictive analytics, and machine learning in the legal sector. This review sets the stage for the subsequent chapters by highlighting the key trends, challenges, and opportunities in these areas. Chapter 3 describes the methodology used, including the data collection process, the development of the predictive model, and the techniques used for feature engineering and model evaluation. The chapter provides a comprehensive overview of the research design, allowing readers to understand how the study was conducted and how the results were obtained. Chapter 4 presents the results, including the feature importance analysis,

the evaluation of the predictive model, and the findings related to the factors that influence bill rejections. Chapter 5 discusses the practical applications of the predictive model and explores the potential for further research and development in this area. This chapter also addresses some limitations of the study and suggests ways in which future research could build on the findings presented here. Finally, Chapter 6 concludes by summarizing the key findings and their implications for the legal industry.

2. Predictive Analysis and Legal Billing

2.1 Introduction to Predictive Analytics in Legal Billing

Predictive analytics is a sophisticated branch of data analysis that leverages historical data, machine learning algorithms, and statistical models to forecast future outcomes (Shmueli et al., 2017). Within the legal industry, the potential for predictive analytics to revolutionize operations, particularly in billing practices, is significant. As legal firms grow in size and scope, the volume of billing data they generate increases exponentially. The management and processing of this data presents challenges that traditional methods -such as manual audits, periodic reviews, and subjective judgment- are increasingly unable to address effectively. The sheer scale and complexity of legal billing data calls for more advanced, data-driven approaches.

In the last decade, interest in the application of predictive analytics to the legal domain has surged (Susskind, 2017). This interest has extended to various aspects of legal practice, including litigation outcome prediction, contract management, and billing processes. Predictive models, which utilize machine learning techniques to identify patterns in large datasets, offer a promising solution to the challenges posed by traditional billing methods. By uncovering relationships and trends that may not be immediately apparent, these models enable firms to anticipate client behaviors, optimize their billing practices, and ultimately minimize the occurrence of bill rejections.

2.2 Key Factors Influencing Bill Rejections

Understanding the factors that influence bill rejections is essential for developing an effective predictive model. The existing industry reports have identified some factors that contribute to bill rejections (Matich, 2019). These factors provided the initial foundation for feature selection and engineering in the predictive model developed in this thesis.

1. **Billing Amount and Rate Structures**: The total amount billed, and the rates charged are critical factors in client decision-making. Bills that deviate significantly from expected amounts or agreed-upon rates are more likely to be rejected or disputed. Clients may question the

justification for higher rates or additional charges, particularly if they are not clearly explained in the billing narrative.

2. **Work Descriptions (Narratives)**: The clarity, specificity, and transparency of work descriptions included in the bill play a significant role in client acceptance. Ambiguous or overly complex narratives can lead to misunderstandings and disputes. For example, a vague description like "Legal research" might prompt a client to question the necessity and extent of the work performed, leading to a rejection or request for further clarification.
3. **Client-Specific Billing Guidelines**: Many clients have specific guidelines regarding how bills should be structured and what costs are acceptable. Failure to adhere to these guidelines is a common cause of bill rejections. These guidelines may include preferred formats, required documentation, and restrictions on billable hours for certain tasks. Predictive models must account for these guidelines to accurately forecast the likelihood of bill rejection.
4. **Timing and Frequency of Billing**: The timing and frequency of billing can significantly impact client acceptance. Bills submitted too late may raise concerns about the accuracy of the recorded hours, while bills submitted too frequently may be perceived as excessive or unnecessary. Clients often have expectations about when they will receive invoices, and deviations from these expectations can lead to disputes.
5. **Project Type and Scope**: The nature of the legal work being billed -whether it involves litigation, corporate law, intellectual property, or other areas- can influence the likelihood of bill acceptance. Different types of legal work are associated with different client expectations and billing standards. For example, corporate clients may have stricter requirements for billing documentation in mergers and acquisitions, while litigation clients may focus more on the outcomes achieved.
6. **Client History and Relationship**: The history of the client-firm relationship, including past disputes or billing issues, can affect the likelihood of bill rejection. Clients who have previously contested bills may be more likely to do so again, particularly if the same issues recur.

Additionally, long-term clients who have developed trust in the firm may be more lenient, while new clients may scrutinize bills more closely.

The identification of these factors allows for the development of a predictive model that is tailored to the specific challenges of legal billing. By incorporating these variables into the model, the research aims to create a tool that can accurately predict bill rejections and provide actionable insights for law firms. While some of these affirmations are consistent with the results obtained, other claims were found at odds with the findings in terms of variable importance.

2.3 Machine Learning Techniques in Predictive Analytics

Despite its potential, recent years have seen limited exploration of predictive analytics in legal billing (Thomson Reuters, 2020). Most existing literature addresses the general application of machine learning techniques across various industries, while relatively few studies examine the specific challenges associated with legal billing (Yu et al., 2022).

Machine learning (ML) forms the backbone of predictive analytics (Breiman, 2001). It involves training algorithms to recognize patterns in data, enabling them to make predictions or decisions without the need for explicit programming for every possible scenario (Hastie et al., 2009). In the context of legal billing, several machine learning techniques are particularly relevant, including regression analysis, decision trees, support vector machines (SVM), and ensemble methods like Random Forests and XGBoost.

Regression Analysis is one of the most basic forms of predictive modeling. It involves identifying the relationship between a dependent variable and one or more independent variables. While powerful in contexts with simple, linear relationships, regression analysis often falls short in complex datasets like those found in legal billing, where interactions between variables are non-linear and multifaceted. This limitation has driven the exploration of more advanced techniques.

Support Vector Machines (SVM) are particularly effective in high-dimensional spaces, where the number of features exceeds the number of samples. SVMs work by finding the

hyperplane that best separates the data into different classes. However, they are computationally expensive and can be challenging to scale for large datasets like those in legal billing, where the number of variables and records can be substantial.

Decision Trees represent another widely used technique. In this approach, data is split into branches based on certain criteria, with each branch representing a decision path leading to a predicted outcome. Decision trees are intuitive and easy to interpret, making them popular in many fields. However, they are prone to overfitting, particularly with complex datasets. To address this, ensemble methods like Random Forests were developed, which combine multiple decision trees to improve predictive accuracy and generalizability[1].

Among these techniques, XGBoost (Extreme Gradient Boosting) has gained significant traction in recent years due to its high performance and efficiency in handling large datasets. XGBoost is an ensemble learning method that builds multiple decision trees sequentially, with each new tree correcting the errors of the previous ones. This approach not only enhances predictive accuracy but also mitigates the risk of overfitting, making it particularly well-suited for the complex and varied data found in legal billing processes (Chen & Guestrin, 2016).

XGBoost's ability to handle missing data and its robustness in the presence of noise, further contribute to its effectiveness in predictive modeling. It has been widely adopted in competitive data science and is recognized for its superior performance in classification tasks. These classification tasks involve predicting a specific category or class label based on input features, for example, determining whether an invoice will be accepted or rejected. This is exactly the type of problem addressed in this thesis, where the goal is to build a model that classifies billing outcomes into binary categories. In the context of legal billing, XGBoost's ability to handle a wide range of feature types, including numerical, categorical, and textual data, makes it a fitting choice for developing predictive models that can quickly and accurately forecast bill rejections (Chen & Guestrin, 2016).

[1] For more on these ML models, see Hastie et al., (2009).

2.4 Expanding on Narrative Analysis in Legal Billing

Narrative analysis is regarded as a critical aspect of legal billing that requires special attention. Legal narratives, which describe the work performed for clients, are inherently complex and varied. They are typically unstructured text data, making them challenging to analyze using traditional methods. However, advancements in natural language processing (NLP) and text mining have opened new avenues for analyzing these narratives in a structured and meaningful way (Kowsari et al., 2019).

In legal billing, narratives must convey the value of the services provided in a clear and concise manner. Clients rely on these descriptions to understand what work was done, why it was necessary, and how it benefited them. Poorly written or vague narratives can lead to confusion, dissatisfaction, and ultimately, bill rejections. For example, a narrative that simply states: "Reviewed documents" provides little insight into the nature or importance of the task, whereas a more detailed description like: "Reviewed and analyzed client contracts for compliance with new regulatory requirements" offers a clearer justification for the time billed.

The application of NLP techniques -such as TF-IDF vectorization- allow for the conversion of unstructured narratives into structured data that can be analyzed using machine learning algorithms. TF-IDF, which stands for Term Frequency-Inverse Document Frequency, is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents (or corpus) (Ramos, 2003). In the context of legal billing, TF-IDF can be used to identify key terms and phrases in billing narratives that are associated with higher or lower rates of bill acceptance.

Additionally, clustering techniques like K-Means can be employed to group similar narratives together, enabling the identification of common themes or patterns in the data. These clusters can then be analyzed to determine which types of narratives are more likely to result in bill rejections.

For instance, narratives related to routine tasks like "drafting" or "research" may be more

prone to rejection if clients perceive them as less valuable or if they lack sufficient detail. By incorporating narrative analysis into the predictive model, the research aims to provide law firms with actionable insights into how their billing descriptions impact client decisions. This, in turn, can help firms improve their billing practices by refining the language used in narratives to better align with client expectations and reduce the likelihood of disputes.

3. Methodology

3.1 Introduction

The primary goal of the research was to develop a predictive model capable of forecasting the likelihood of bill rejections based on historical billing data. To achieve this, a systematic approach encompassing data collection, data preprocessing, feature engineering, model development -using the XGBoost algorithm-, and evaluation of the model's performance was employed.

3.2 Data Collection

The collected data forms the foundation for building the predictive model. This thesis uses a dataset drawn from the aforementioned law firm's historical billing records, covering a one-year period and capturing a diverse range of client interactions, billing practices, and project details. The dataset consisted of over 160,000 entries, each representing a billable event[2], with attributes such as:

- **Billing Amounts:** This includes the total amount billed, standard rates, and project- specific rates, crucial for understanding how pricing influences rejections.
- **Project Details:** These include information about the type of project, the legal area it pertains to, and a broad description of the type of work carried out.
- **Employee Information:** This captures details about the employees responsible for the work, including their position, group affiliation, geographical location, and area of expertise.
- **Work Narratives:** These are textual descriptions that detail the work performed. Narratives are often critical in client evaluations, as they provide context and justification for the billed hours.
- **Billing Outcomes:** The dataset also records whether the bill was contested or if a discount had to be given. These discounts served as the target variable for the predictive model.

Despite coming from a single law firm, the dataset's diversity in terms of clients, projects, and

[2] See Appendix A for a complete list of the dataset's variables.

outcomes provided a rich source of data, so that the predictive model developed is a robust and generalizable one.

Exploratory data analysis (EDA) played a crucial role in this study by laying the foundation for understanding the underlying features and interactions within the dataset. Despite being collected cautiously and with great care, billing data had many variables with a high count of missing values before data preprocessing was done, as shown in Figure 1. EDA facilitated identifying key patterns and correlations between variables -such as seasonal fluctuations illustrated in Figure 2- which informed the subsequent steps in the data preprocessing phase. This analysis highlighted potential outliers but also revealed complex relationships between different features, such as the interplay between billing amounts, project types, and narrative content.

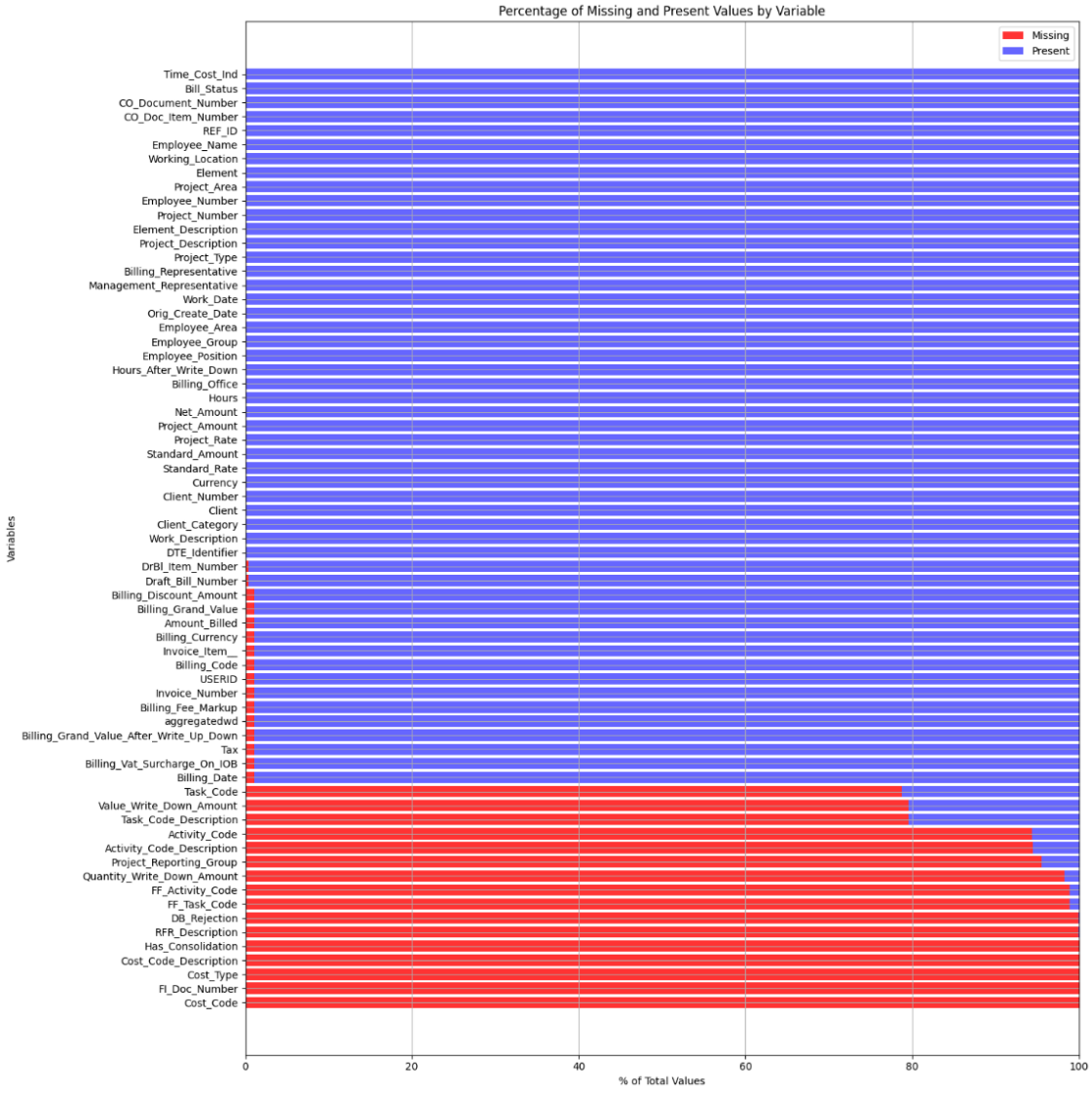

Figure 1 - Percentage of Missing and Present Values by Variable (Before Pre-processing)

Gaining these insights early on allowed for the design of an effective preprocessing strategy, ensuring that the predictive model could be built on a robust and well-prepared dataset.

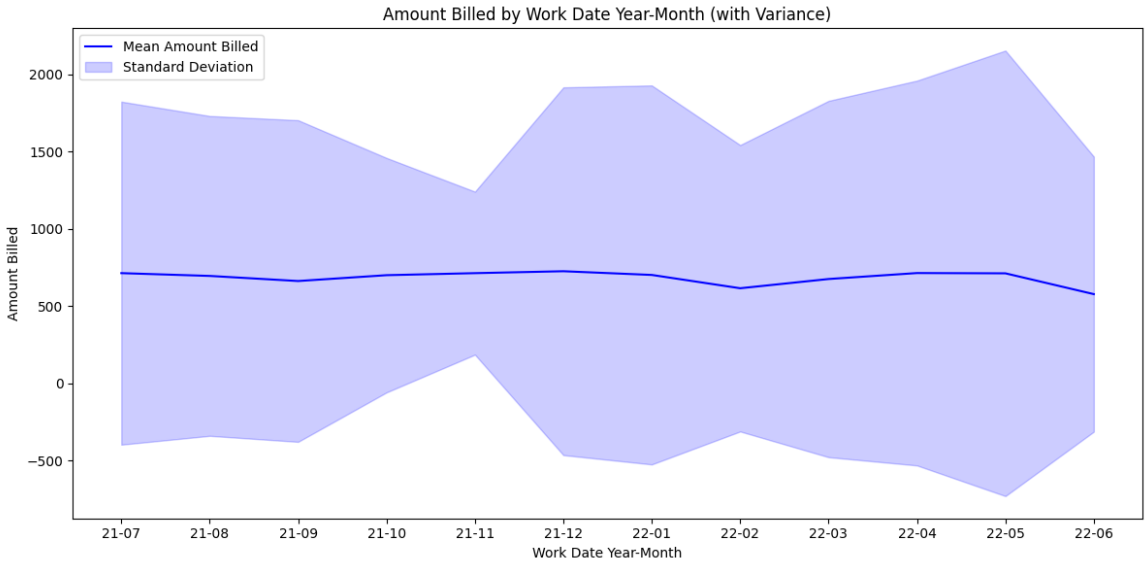

Figure 2 - Amount Billed by Work Date Year-Month (with Variance)

3.3 Data Preprocessing

Data preprocessing is a critical step that prepares the raw data for analysis and modeling. Given the complexity and volume of the dataset, several preprocessing steps were necessary to ensure data quality and consistency, thereby enhancing the performance of the machine learning model.

3.3.1 Data Cleaning

Data cleaning was the first step in preprocessing and involved addressing inconsistencies, missing values, and irrelevant features.

- **Missing Values:** Missing data in numeric columns was replaced with 0 when appropriate, with the column mean to prevent skewing, or their rows were removed entirely from the dataset if deemed unnecessary or distortive for the analysis -like with Cost Code, as they represent costs and not bills-. Figure 3 shows the percentage of missing values for the variables kept on the dataset and the ones added after some preprocessing, but before this imputation.
- **Irrelevant Features:** Features that did not contribute to the predictive task or that could lead to

data leakage, such as unique identifiers like "*Invoice Number*" and "*Employee Name*" were removed. This decision was based on exploratory data analysis and was further validated by feature importance assessments later in the modeling process. Using Correlation Analysis, vital interactions between the variables were uncovered, as can be seen in the correlation matrix shown in Figure 4. This allowed for better feature selection and engineering.

- **Handling Outliers:** Outliers, particularly in billing amounts and hours, were intentionally not modified or removed. This decision was made to retain the influence of "extreme" values, as these higher or lower amounts might provide valuable insights into patterns of billing discounts or rejections. By preserving these outliers, the analysis aimed to capture any potential correlations between extreme values and the likelihood of discounts or rejections.

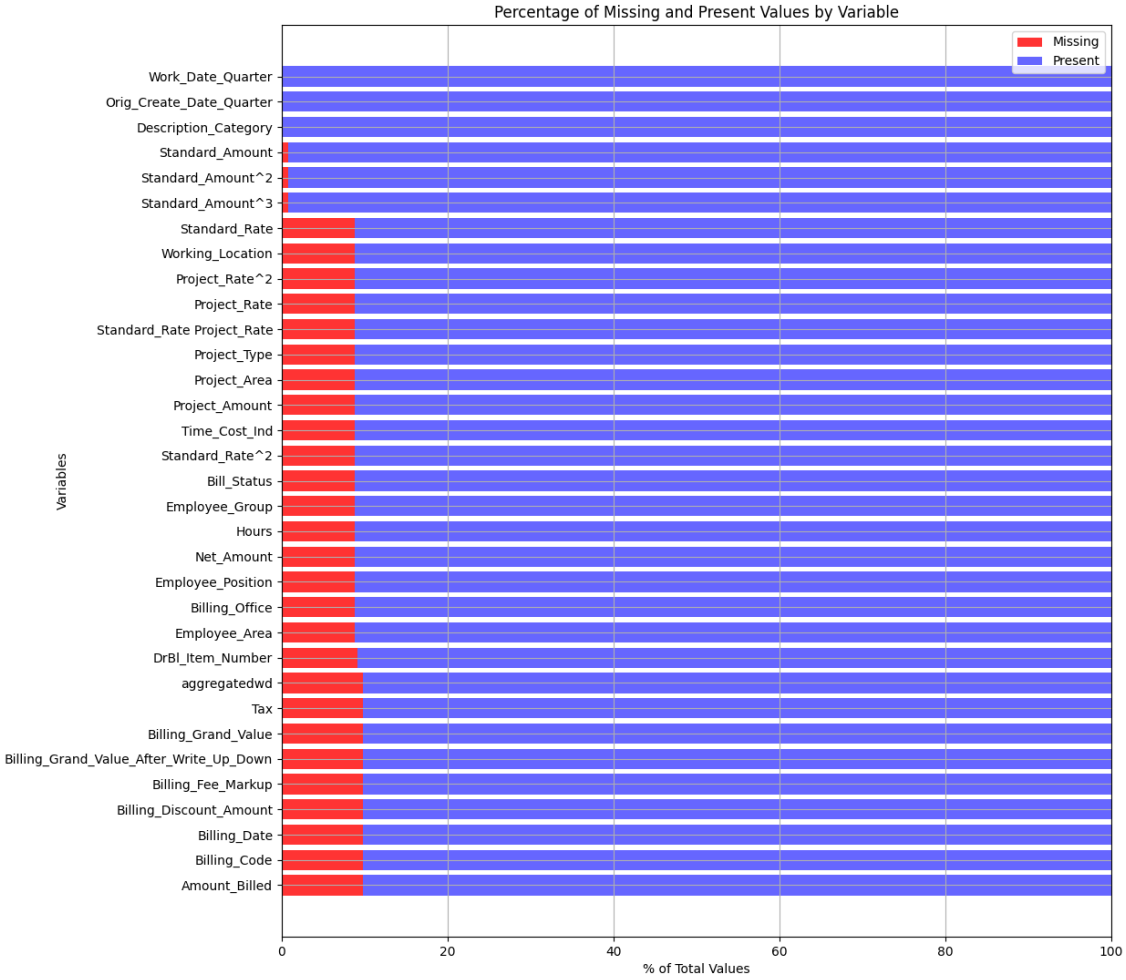

Figure 3 - Percentage of Missing and Present Values by Variable (Before Imputation)

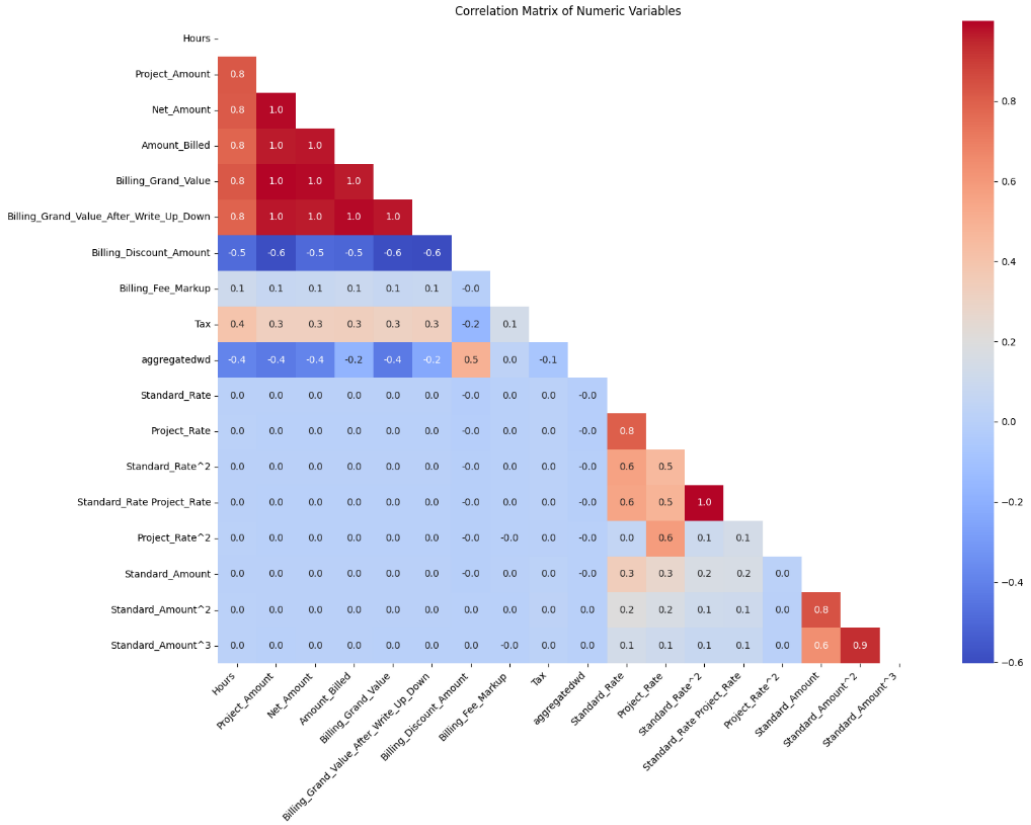

Figure 4 – Correlation Matrix for Numeric Variables

3.3.2 Feature Engineering

Feature engineering involved creating new variables from the existing data to enhance the predictive power of the model. This process enabled the machine learning algorithm to capture complex patterns and relationships within the data.

- **Polynomial Features:** Polynomial transformations were applied to certain numerical variables such as "*Standard Rate*" and "*Project Rate*". These transformations allowed the model to capture non-linear relationships between these rates and the likelihood of bill rejection. For example, adding second-degree polynomial features helped model interactions between these rates, which might indicate billing anomalies.
- **Date Transformations:** The process also involved converting dates related to billing and work completion into categorical variables representing quarters ("*Work Date Quarter*", "*Orig Create Date Quarter*"). This transformation allowed the model to better capture seasonal patterns and trends in billing practices and client behavior, which could influence bill rejections.

- **Aggregated Write-Down Amount:** The difference between the amount billed and the project amount after accounting for write-downs became a new feature, "*aggregatedwd*". Though this feature was originally present in the dataset as "*Value Write-Down Amount*", around 80% of the rows were missing a value. After re-creating it by combining other features, most of the entries had this feature, crucial for understanding the dataset through exploratory data analysis. This feature had a perfect -1 correlation with "*Value Write-Down Amount*".

3.3.3 Narrative Vectorization

A unique challenge in this analysis was handling the work narratives, which are textual descriptions of the tasks performed. These narratives were vectorized using the Term Frequency-Inverse Document Frequency (TF-IDF) method (Ramos, 2003). TF-IDF is a statistical measure that reflects the importance of a word in a document relative to its frequency across a collection of documents.

- **Tokenization:** The process involved tokenizing narratives, breaking down the text into individual words or tokens. Common *stop words* (e.g., "and" "the") were excluded, as they do not contribute meaningful information to the analysis.
- **TF-IDF Calculation:** Each token is assigned a weight based on its frequency in the narrative relative to its frequency across all narratives in the dataset. This weighting process emphasized terms that were unique to specific narratives, helping the model differentiate between dissimilar types of work descriptions.
- **Clustering with K-Means:** Following vectorization, the narratives are clustered using the K-Means algorithm into 15 categories. Each narrative is then labeled with the most representative cluster, transforming the complex text data into a categorical variable that could be used as an input to the machine learning model.

  To validate the approach, the study involved exploring an alternative method using a

Sentence Transformer model -*all-MiniLM-L6-v2*[3]-, a deep learning-based approach for generating sentence embeddings. Sentence Transformer models convert entire sentences into numerical vectors by capturing their semantic meaning, rather than simply analyzing word frequency. This approach is particularly useful when the goal is to preserve contextual nuances, such as word order and meaning, which can be important in understanding legal narratives. However, while this technique has advantages in capturing complex relationships between words, it also presents important limitations within the scope of this research. Sentence Transformer models typically require more memory and processing power, leading to considerably longer execution times, which could hinder usability in environments without advanced hardware or cloud infrastructure.

Given that the primary goal of this thesis is to provide a lightweight, transparent, and efficient predictive tool that could be implemented in real-world legal settings, TF-IDF was ultimately deemed the more appropriate choice. Unlike Sentence Transformers, TF-IDF allows for straightforward interpretation by ranking terms based on their distinctiveness, enabling the resulting narrative clusters to be more easily explained using identifiable keyword groupings. This added layer of interpretability is particularly valuable when presenting results to legal professionals who may not have technical backgrounds but need clear insight into the rationale behind model predictions. In this context, the clarity, speed, and lower computational demand of TF-IDF made it better aligned with the practical objectives of the project.

Out of the original 66 variables in the raw dataset, a total of 41 were dropped during preprocessing and feature engineering. This reduction focused on removing features that were irrelevant, highly incomplete, or likely to introduce data leakage. To enhance the model's predictive performance, 8 new variables were engineered, including polynomial transformations of numeric features, categorical representations of date fields, and narrative cluster labels from K-Means analysis of the work descriptions. The final analytic dataset consisted of 33 variables: 25 retained from the original and 8 newly engineered, plus the target and row identifier.

[3] The model can be found here: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

3.4 Model Development

The core of this thesis is the development of a predictive model using the XGBoost algorithm. XGBoost was selected due to its robustness, efficiency, and ability to handle large, complex datasets with numerous features. The model development process involved several key steps:

3.4.1 Data Splitting

The workflow involved dividing the dataset into three subsets to ensure that the model was trained, validated, and tested effectively:

- **Training Set (70%):** This subset allowed the model to learn by adjusting its parameters to minimize prediction errors.
- **Validation Set (15%):** This subset supported the model's hyperparameter tuning and overfitting prevention by evaluating its performance on data not seen during training.
- **Test Set (15%):** This final subset allowed the evaluation of the model's performance, simulating its behavior on entirely new data in a real-world setting.

This splitting ensured that the model could generalize well to new data, avoiding overfitting while maximizing predictive accuracy.

3.4.2 Model Training

The XGBoost model was trained using the prepared dataset, ensuring a balance between model complexity, generalizability, and performance optimization. Given the structured nature of the legal billing data, XGBoost's ability to handle categorical variables directly was leveraged, eliminating the need for extensive preprocessing.

The training process incorporated early stopping to mitigate overfitting and enhance model stability. The model monitored validation performance and halted training if performance did not improve after a set number of iterations, preventing unnecessary complexity. Additionally, the evaluation metric was set to AUC-PR (Area Under the Precision-Recall Curve) to ensure that the model prioritized identifying rejected bills accurately while minimizing false positives.

One of the key challenges in model training was addressing class imbalance, as rejected bills represented a minority of the total observations. The approach involved applying two different techniques to explore solutions for this issue: the scale_pos_weight hyperparameter in XGBoost and Synthetic Minority Over-sampling Technique (SMOTE). The scale_pos_weight method adjusts the importance of the minority class by weighing its contribution to the loss function, while SMOTE creates synthetic examples of the minority class to balance the dataset. However, when the dataset is fully balanced using SMOTE, the impact of scale_pos_weight becomes irrelevant, as the class distribution is already equalized.

The process involved using the validation set to assess model performance, with a focus on achieving and maintaining a meaningful balance between recall and precision. This strategy ensures reliable detection of rejected bills without generating excessive false positives or introducing bias toward either class.

3.4.3 Hyperparameter Tuning

Hyperparameter tuning was a critical step in optimizing the XGBoost model's performance. Instead of manually selecting hyperparameters, a random grid search approach allowed exploring a range of potential configurations, making the fine-tuning of the model more efficient. The following hyperparameters were selected for tuning based on their impact on model performance:

- **Number of Boosting Rounds (nrounds):** Set between 50 and 200 iterations. A larger number of boosting rounds allows the model to refine its predictions, but too many rounds can lead to overfitting. Early stopping was used to determine the optimal stopping point dynamically.
- **Max Depth:** Varied between 5 and 10**.** This parameter controls the maximum depth of each decision tree in the model. Deeper trees can capture complex interactions in the data, but they also risk overfitting. The chosen range allowed for a trade-off between complexity and generalizability.
- **Learning Rate (eta):** Set between 0.1 and 0.3. A lower learning rate ensures that the model learns more gradually, preventing large, unstable updates to weights, which can cause

fluctuations in performance.

- **Gamma:** Tuned between 0.0 and 0.5. Gamma acts as a regularization parameter, determining the minimum loss reduction required before a node split occurs. A higher gamma value makes the model more conservative, helping to reduce overfitting.
- **Colsample_bytree:** Ranged from 0.3 to 1.0. This parameter controls the fraction of features randomly selected for training each tree, introducing diversity among the trees and reducing the risk of overfitting.
- **Min_child_weight:** Set between 2 and 10. This parameter specifies the minimum sum of instance weights required in a leaf node. A higher value results in more generalized trees by preventing small, overly specific splits.
- **Subsample:** Varied between 0.3 and 1.0. This parameter determines the fraction of training samples used for each tree. A lower subsample value reduces overfitting by ensuring that trees are not overly specialized to particular subsets of the data.
- **Scale_pos_weight:** Dynamically computed as the ratio of the majority to minority class instances but only applied in unbalanced datasets. In the balanced dataset, this parameter had no effect since the class distribution was already evened out.

By using random grid search, the tuning process efficiently explored multiple combinations of these hyperparameters. The selection criteria prioritized AUC-PR (Area Under the Precision-Recall Curve) over accuracy, ensuring the model effectively balanced recall and precision, critical for correctly identifying rejected bills. The final configuration was determined based on the best validation performance, ensuring that the model generalized well to unseen data.

The goal of this thesis is to build a lightweight, explainable tool that can be implemented in legal environments. Random grid search matches this approach by being straightforward, interpretable, and robust, especially when paired with early stopping and validation set monitoring. While more advanced optimization libraries such as Optuna[4] could offer efficiency gains through

[4] For information on Optuna visit: https://optuna.org/

adaptive sampling and pruning strategies, random grid search was chosen for its simplicity, transparency, and alignment with the project's objectives. This method provides full visibility into the parameter combinations being tested, which is particularly valuable in professional and academic contexts where clarity and reproducibility are essential. It also allows for direct control over the range and distribution of each hyperparameter, making it easier to incorporate domain-specific knowledge. Furthermore, the strong model performance achieved using random grid search indicated that more complex optimization strategies were not required to meet the goals of this research.

3.4.4 Why XGBoost?

Selecting the right machine learning model for predicting bill rejections involved assessing various alternatives. XGBoost (Extreme Gradient Boosting) emerged as the preferred option due to its superior performance, interpretability, and scalability, particularly in handling structured legal billing data with a mix of numerical, categorical, and text-derived features.

Traditional regression models, such as logistic regression, are often a great solution for simpler problems due to their ease of use and interpretability. However, these models assume a linear relationship between features and outcomes, which does not align well with the non-linear nature of legal billing decisions. Additionally, they may struggle with multicollinearity between financial variables like "Net Amount" and "Project Amount," potentially leading to unstable predictions. While polynomial feature engineering could help capture interactions, this approach becomes computationally impractical for large datasets. In this dataset, the logistic regression model demonstrated predictive performance equivalent to random chance, offering no improvement over simply guessing outcomes.

Support Vector Machines (SVMs) could also be a good alternative for their ability to classify complex data. However, SVMs require substantial computational resources and exhibit slow loading and prediction times, even when using a linear kernel. This limitation is especially pronounced when processing large datasets or managing a high number of features, as was the

case for a dataset of this size, which exceeds 160,000 billable entries. The model needs to compute distances between data points and execute complex kernel functions, demanding substantial processing resources and increasing runtime. Given the objective of delivering a lightweight, responsive prediction tool for law firms, this prolonged execution time makes SVM unsuitable for the intended application. Furthermore, SVMs require extensive preprocessing to handle categorical variables whereas XGBoost can process them directly. They also offer limited feature interpretability, making it difficult to extract actionable insights from predictions,

Random Forests, another ensemble method, did not achieve the same level of performance as XGBoost in this study. While Random Forests are known for their robustness and ability to handle a wide variety of data types, they exhibited higher memory consumption and slower inference speed, making them less practical for large-scale or real-time applications. Additionally, because Random Forests build each tree independently and aggregate their predictions through majority voting, they lack the iterative error-correcting mechanism that characterizes XGBoost's sequential boosting approach. As a result, XGBoost can progressively correct for mistakes made by previous trees, which leads to higher overall accuracy. Furthermore, the independent nature of tree construction in Random Forests can miss subtle patterns in the data that XGBoost is able to capture through its gradient boosting framework.

Consequently, XGBoost ultimately stood out for its ability to handle imbalanced data, a key concern given that rejected bills make up a minority of the dataset. Its scale_pos_weight parameter helped balance the classes, improving recall without significantly reducing precision. While SHAP values can also be used with other models like Random Forests to interpret feature importance, XGBoost integrates more efficiently with SHAP thanks to better internal compatibility. As a result, SHAP values from XGBoost tend to be more stable across runs and easier to visualize. This interpretability is especially valuable in legal billing, where understanding the reasons behind bill rejections is just as important as predicting them accurately.

Beyond performance, efficiency was critical for real-world application. XGBoost is highly

optimized for speed, supporting parallel computation and efficient memory usage, making it significantly faster than Random Forests and SVMs for both training and inference. This is especially relevant, considering part of the goal of this paper was providing a fast and effective tool to allow non-technical users to predict billing rejections. XGBoost also handles missing values natively, unlike logistic regression, which requires imputation strategies that can introduce bias.

In summary, XGBoost delivers higher accuracy, faster processing, better interpretability, and improved handling of imbalanced data compared to alternative methods. Its ability to combine predictive power with actionable insights made it the best choice for forecasting bill rejections in legal billing.

3.5 Model Evaluation

The final model was rigorously evaluated on the test set to determine its predictive accuracy and robustness. Several metrics were used to evaluate the model's performance:

- **Accuracy:** Accuracy measures the proportion of correct predictions out of the total number of cases. It provides a general measure of the model's overall performance.
- **Precision and Recall:** Precision measures the proportion of positive predictions that are actually correct, while recall measures the proportion of actual positives that are correctly identified. These metrics are crucial in understanding the trade-offs between predicting too many false positives versus missing true positives.
- **AUC-ROC (Area Under the Receiver Operating Characteristic Curve):** The AUC- ROC score provides a single measure of model performance by evaluating the trade-offs between true positive rates and false positive rates across different threshold settings. A higher AUC indicates a better-performing model.
- **SHAP Values:** SHapley Additive exPlanations (SHAP) values help to explain the contributions of each feature to the model's predictions. This analysis helped in understanding which features were most influential in predicting bill rejections, offering insights into the decision-making process of the model.

4. Results

4.1 Overview of Results

The results of this study center around the performance of the predictive model in forecasting bill rejections, the relative importance of different features used in the model, and the specific role of narrative content in influencing bill acceptance or rejection. This section presents a detailed analysis of the model's performance, the key factors contributing to its predictions, and the practical implications of these findings for legal billing practices. The results validate the effectiveness of using predictive analytics in legal billing and provide insights into how various features, including narrative content, influence billing outcomes.

4.2 Model Performance Metrics

The predictive model was trained under three different dataset conditions: (1) an unbalanced dataset without using the XGBoost *"scale_pos_weight"* hyperparameter, (2) an unbalanced dataset utilizing *"scale_pos_weight"* to account for class imbalance, and (3) a balanced dataset obtained through SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic samples of the minority class. The model's performance was then assessed on a validation and ultimately a test set using accuracy, precision, recall, and AUC-ROC. These metrics provide insight into how well the model distinguishes between accepted and rejected bills across these different training conditions.

Understanding Model Performance

To better illustrate how the model performed in each case, confusion matrices were generated for both the validation and test sets across the three configurations. Figure 5 shows the Confusion Matrices for the model in both the validation and test set. It indicates both the number of True Positives and True Negatives the model was able to predict, and the number of False Positives and False Negatives the model labeled erroneously. This breakdown provides a comprehensive view of the model's classification performance in practical terms for each of the 3 dataset conditions mentioned above.

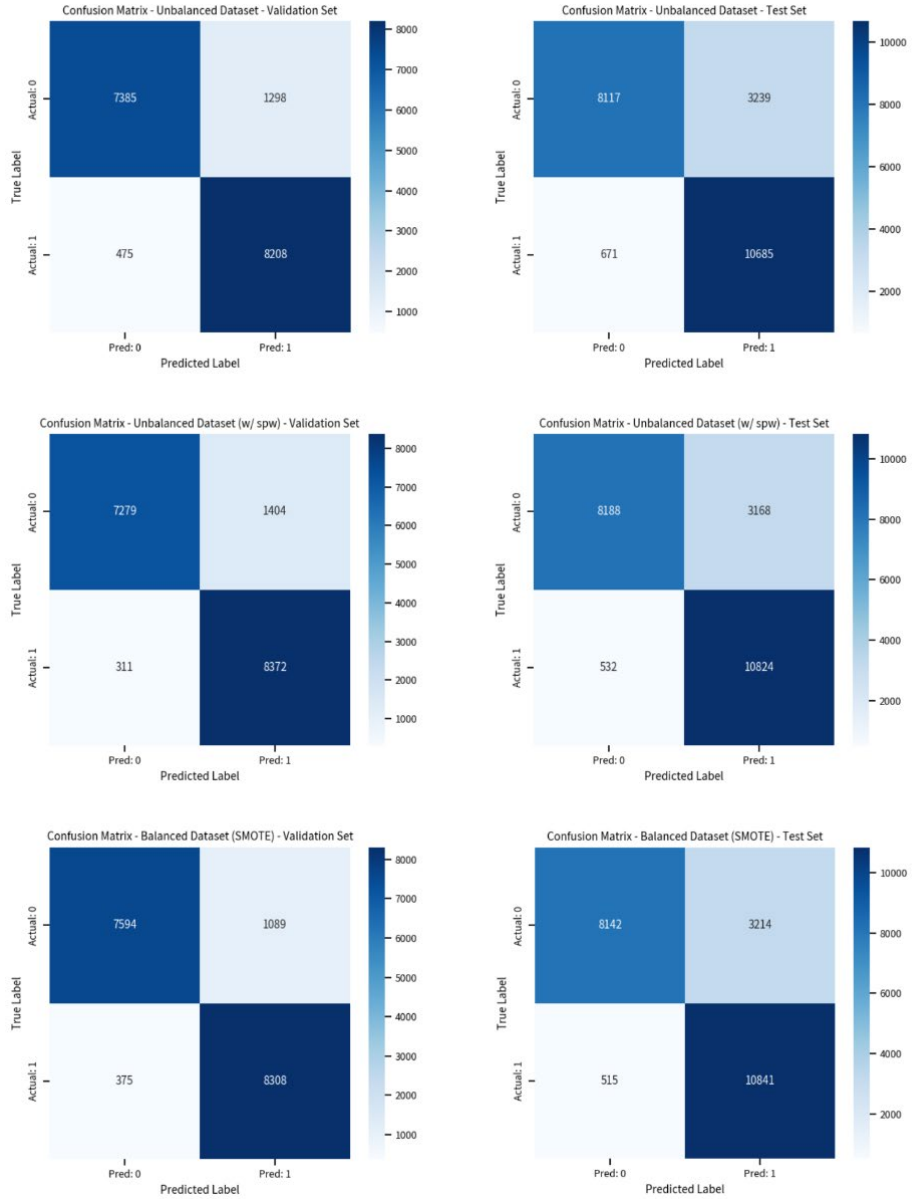

Figure 5 – Confusion Matrices Showing the Performance on Validation and Test Sets for the Models Trained Using an Unbalanced Dataset, an Unbalanced Dataset Using the scale_pos_weight Hyperparameter, and a Balanced Dataset Using SMOTE

Table 1 contains the components of the Confusion Matrix across the 3 dataset conditions for the Validation Set.

Validation Set - Confusion Matrix Components

Setup	True Positives (TP)	False Positives (FP)	False Negatives (FN)	True Negatives (TN)
Unbalanced (No scale_pos_weight) - Validation	8208	1298	475	7385
Unbalanced (With scale_pos_weight) - Validation	8372	1404	311	7279
Balanced (Oversampled Minority Class) - Validation	8308	1089	375	7594

Table 1 – Confusion Matrix Components for the Models Trained Using an Unbalanced Dataset, an Unbalanced Dataset Using the scale_pos_weight Hyperparameter, and a Balanced Dataset Using SMOTE on the Validation Set

Table 2 contains the components of the Confusion Matrix across the 3 dataset conditions for the Test Set.

Test Set - Confusion Matrix Components

Setup	True Positives (TP)	False Positives (FP)	False Negatives (FN)	True Negatives (TN)
Unbalanced (No scale_pos_weight) - Test	10685	3239	671	8117
Unbalanced (With scale_pos_weight) - Test	10824	3168	532	8188
Balanced (Oversampled Minority Class) - Test	10841	3214	515	8142

Table 2 – Confusion Matrix Components for the Models Trained Using an Unbalanced Dataset, an Unbalanced Dataset Using the scale_pos_weight Hyperparameter, and a Balanced Dataset Using SMOTE on the Test Set

The confusion matrices reveal a minor shift in predictive performance when transitioning from an unbalanced dataset to a balanced one. The introduction of scale_pos_weight in the unbalanced dataset leads to marginal improvements, particularly through a slight reduction in false negatives. In contrast, balancing the dataset through synthetic oversampling results in a more noticeable decrease in false negatives and a corresponding increase in overall accuracy.

Key Performance Metrics Comparison

Each approach resulted in different performance trade-offs in terms of performance, particularly in how well the model balanced precision and recall. Table 3 shows the evaluation metrics for the Validation Set.

Validation Set - Evaluation Metrics

Setup	Precision	Recall	Accuracy	ROC AUC	F1 Score
Unbalanced (No scale_pos_weight) - Validation	0.863	0.945	0.897	0.948	0.902
Unbalanced (With scale_pos_weight) - Validation	0.856	0.964	0.901	0.95	0.906
Balanced (Oversampled Minority Class) - Validation	0.884	0.956	0.915	0.962	0.918

Table 3 – Evaluation Metrics for the Models Trained Using an Unbalanced Dataset, an Unbalanced Dataset Using the scale_pos_weight Hyperparameter, and a Balanced Dataset Using SMOTE on the Test Set

Table 4 shows the evaluation metrics for the Test Set.

Test Set - Evaluation Metrics

Setup	Precision	Recall	Accuracy	ROC AUC	F1 Score
Unbalanced (No scale_pos_weight) - Test	0.767	0.94	0.827	0.833	0.845
Unbalanced (With scale_pos_weight) - Test	0.773	0.953	0.837	0.843	0.853
Balanced (Oversampled Minority Class) - Test	0.771	0.954	0.835	0.852	0.853

Table 4 – Evaluation Metrics for the Models Trained Using an Unbalanced Dataset, an Unbalanced Dataset Using the scale_pos_weight Hyperparameter, and a Balanced Dataset Using SMOTE on the Test Set

- **Precision Across Setups:** Precision remains consistently high across all three scenarios, indicating that the model maintains strong performance in minimizing false positives regardless of whether the dataset is balanced or not. This means the model is reliably predicting rejected bills with a low rate of incorrectly labeling acceptable ones.
- **Impact of Balancing on Recall:** The most significant gain in performance appears in recall, which improves from 0.940 in the unbalanced (no scale_pos_weight) setup to 0.954 in the balanced dataset. This indicates that the model trained on balanced data is more effective at capturing true positives. Effectively capturing rejected bills helps minimize the risk of undetected rejections, which is critical in a legal billing context where they can translate into substantial financial losses and reputational damage.
- **Overall Accuracy Trends:** Accuracy can be slightly higher in the unbalanced setups, expected given that the majority class dominates in an imbalanced dataset, but that is not observed. Accuracy alone is not a sufficient measure of model performance in this context, especially when the goal is to correctly flag minority class instances (rejected bills). The SMOTE model sacrifices a small fraction of accuracy (still higher than the unbalanced at 0.835) in exchange for a major gain in recall, which is a more valuable trade-off.
- **F1 Score as a Balanced Metric:** The F1 Score, which balances precision and recall, shows a progressive improvement with better handling of class imbalance. It increases from 0.845 in the first setup to 0.853 in the balanced dataset, confirming that the overall quality of the model's predictions improves significantly with balancing.
- **Consistency in ROC AUC:** The ROC AUC remains very high across all setups, ranging from 0.833 to 0.852. This shows that all models perform well in distinguishing between the two classes. However, the highest AUC is achieved with the balanced dataset, which aligns with the idea that a more representative training distribution leads to better generalization.

Threshold Optimization and ROC Curve Analysis

To maximize predictive performance, an optimal threshold was determined for each scenario. Table 5 shows the optimal thresholds for the 3 dataset conditions.

Test Set - Best Thresholds

Setup	Best Threshold
Unbalanced (No scale_pos_weight)	0.316
Unbalanced (With scale_pos_weight)	0.388
Balanced (Oversampled Minority Class)	0.266

Table 5 – Optimal Thresholds for the Models Trained Using an Unbalanced Dataset, an Unbalanced Dataset Using the scale_pos_weight Hyperparameter, and a Balanced Dataset Using SMOTE on the Test Set

At the optimal classification thresholds, identified using the Youden Index to balance sensitivity and specificity, the balanced dataset consistently exhibited the most favorable trade-off between precision and recall. This reinforces the effectiveness of dataset balancing as a strategy for improving the model's ability to correctly identify rejected bills while minimizing false positives. Figure 6 shows the ROC curve for the Balanced Dataset.

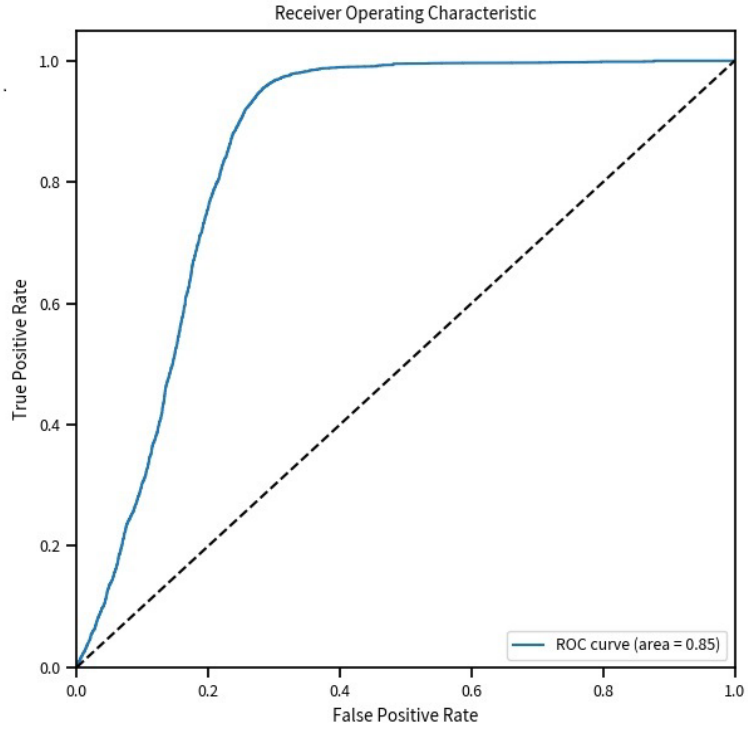

Figure 6 – ROC Curve for the Balanced Dataset

The ROC curve highlights that, across all configurations, the model maintains strong performance in distinguishing between accepted and rejected bills. However, the curve for the balanced dataset shows a slightly steeper incline, indicating better discrimination.

Precision-Recall Curve for Balanced Data

Different thresholds affect the trade-off between true positives and false positives. The Precision-Recall Curve plot is useful for evaluating model performance -even more so when dealing with imbalanced datasets-, as bill rejections may be less frequent than acceptances. This is particularly important for the problem we are addressing, as doing additional work to review a suspected entry is less costly than getting a bill rejected by a client or being asked for a discount, which can strain an ongoing relationship. Figure 7 shows the Precision-Recall Curve for the Balanced Dataset.

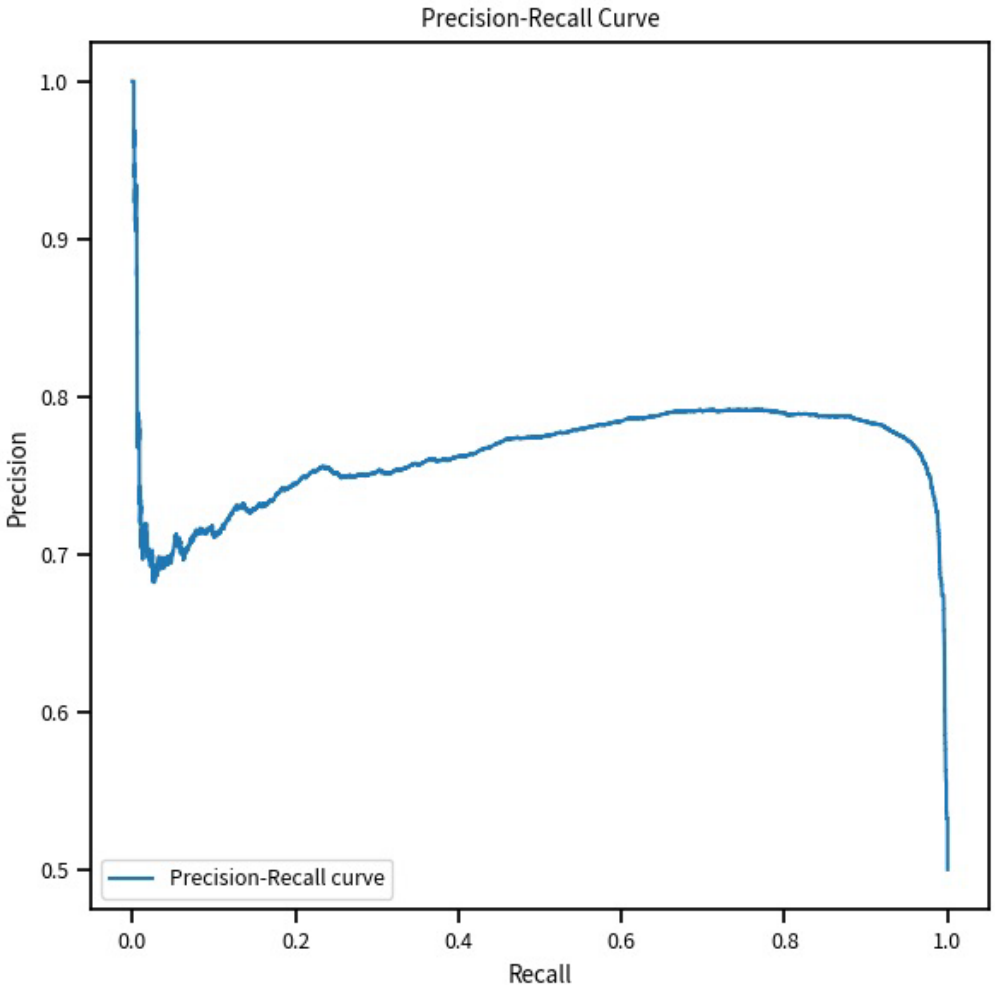

Figure 7 – Precision-Recall Curve for the Balanced dataset

The Precision-Recall curve further confirms the advantages of balancing the dataset. The curve for the balanced dataset is significantly higher, indicating that at most threshold levels, the model retains a better balance between precision and recall.

Final Takeaways

The comparison of different training approaches suggests that balancing the dataset through SMOTE leads to the most significant performance improvements. While using scale_pos_weight on an unbalanced dataset provides a slight boost in recall, it does not result in a drastic change in performance. In contrast, synthetic oversampling the minority class results in the highest accuracy, recall, and precision, making it the most effective strategy.

Maximizing recall is crucial in the context of legal billing because missing a rejected bill can lead directly to financial losses, strained client relationships, or even reputational damage for the firm. In contrast, flagging a bill that would not have been rejected simply results in a second review or minor improvements, which, while slightly inefficient, is far less costly than missing a real issue. Therefore, prioritizing recall ensures that as many potential rejections as possible are captured, safeguarding revenue and maintaining trust with clients.

Comparative Model Performance

To contextualize the performance of the XGBoost model, results from Logistic Regression and Random Forests models served as benchmarks. As shown in the evaluation metrics for the test set contained in Table 6, XGBoost achieved an accuracy of 0.835, a precision of 0.771, a recall of 0.954, and an AUC-ROC of 0.852 when trained on a balanced dataset with oversampled minority class. In contrast, Logistic Regression delivered substantially lower results, with an accuracy of 0.502 and an AUC-ROC of just 0.506, performance nearly equivalent to random guessing. Random Forest showed modest improvement over Logistic Regression, achieving an accuracy of 0.552 and an AUC-ROC of 0.546, but still lagged behind XGBoost in every metric. These results showcase XGBoost's ability to distinguish between accepted and rejected bills in this setting, offering a more reliable and practical solution for predictive legal billing.

Test Set - Model Evaluation Metrics

Model	Precision	Recall	Accuracy	ROC AUC	F1 Score
Logistic Regression	0.501	0.915	0.502	0.506	0.647
Random Forests	0.545	0.618	0.552	0.546	0.578
XGBoost	0.771	0.954	0.835	0.852	0.853

Table 6 - Evaluation Metrics for the Logistic Regression, Random Forests and XGBoost Models on the Test Set

The test set evaluation highlights the distinct advantages offered by XGBoost in this legal billing context. Its high recall indicates a robust ability to identify true bill rejections, reducing the likelihood of overlooked negative outcomes. Precision also remained consistently strong, reflecting a low incidence of false alarms and ensuring that accepted bills were seldom mislabeled as rejections. By contrast, Logistic Regression, despite its efficiency on smaller or simpler datasets, failed to capture the complex patterns present in this data, resulting in near-random classification performance. Random Forests, although an improvement over Logistic Regression, did not reach the threshold of accuracy or reliability required for operational deployment in a legal setting. The comprehensive superiority of XGBoost across all major metrics reinforces its suitability for applications where both accuracy and interpretability are essential.

4.3 Feature Importance Analysis

One of the key advantages of using XGBoost is its ability to provide insights into the importance of different features in making predictions. The SHAP feature importance analysis identifies the 20 most relevant features in determining bill rejection likelihood, highlighting which factors most significantly contribute to the model's decision-making process. By understanding the relative influence of each feature, stakeholders can better interpret model outcomes and gain actionable knowledge about the drivers behind bill rejections, ultimately supporting more informed decision-making in billing practices.

Figure 8 shows the SHAP Summary Plot for the variables influencing the model.

Interpretation of the SHAP Visualization: The SHAP summary plot provides an intuitive way to understand how individual features contribute to prediction outcomes:

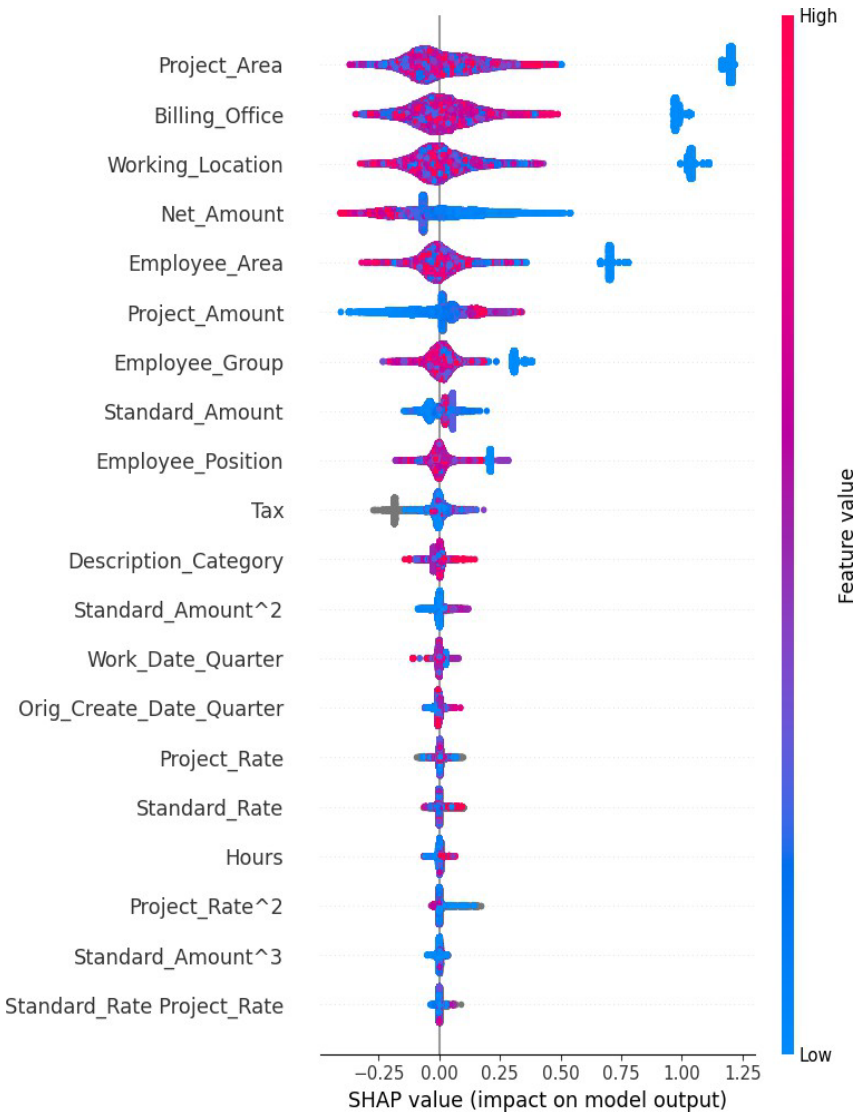

Figure 8 – SHAP Summary Plot

- Each dot represents a single invoice, with color indicating the feature value (red for high, blue for low).
- The x-axis represents SHAP values, where negative values decrease rejection probability, and positive values increase it.
- The vertical spread of dots -y-axis- within each feature represents how much its impact varies across different invoices. Wider spreads indicate more variability in how the feature influences rejection probability, while tightly clustered points suggest more consistency.

- Features with a wide horizontal spread (e.g., "*Net Amount*", "*Project Area*", "*Billing Office*") indicate a strong but possibly varying contribution to rejection predictions, meaning their influence is observed across multiple invoices.
- Features with narrower spreads (e.g., "*Project Rate²*", "*Standard Amount²*") have more localized effects, meaning they are influential in specific cases but do not broadly impact overall billing outcomes.

Key Influential Features: The "*Net Amount*" emerges as the most dominant numerical factor affecting billing outcomes, as indicated by the wide horizontal spread of SHAP values. The plot illustrates that higher net amounts (red-colored dots) systematically increase rejection probability, pushing SHAP values toward the right. Conversely, lower net amounts (blue dots) reduce rejection likelihood, reinforcing the idea that larger invoices attract more scrutiny and are more likely to require justification. This financial influence is also evident for "*Project Amount*", which follows a similar pattern: projects with higher billed amounts see greater rejection risk, likely due to stricter review processes for costly projects.

Among categorical features, "*Project Area*", "*Billing Office*", and "*Working Location*" stand out as major drivers of rejection probability. The horizontal dispersion of SHAP values for these features suggests that certain project areas, billing offices, and working locations have historically been associated with a higher likelihood of rejection, while others contribute to a lower probability. The clustering of high SHAP values implies consistent risk patterns for certain categories within these features, emphasizing the need to identify and address potential regional billing discrepancies.

Employee-related factors, including Employee Area, Employee Group, and Employee Position, exhibit moderate influence on billing decisions. Their SHAP distributions indicate both positive and negative contributions to rejection likelihood, suggesting that certain roles or departments are linked to increased discounts. This variability may arise from differences in project types, documentation standards, or negotiation styles of specific employee groups.

Categorical vs. Numerical Feature Influence: The categorical variables in this model, namely, "*Project Area", "Billing Office", "Working Location", "Employee Area", "Employee Group", "Employee Position", "Work Date Quarter", "Orig Create Date Quarter", and "Description Category"* do not influence rejection likelihood through magnitude but rather through distinct categories with different risk levels. Instead of assuming a linear relationship, the model identifies specific categories that consistently push the likelihood of bill rejection up or down.

- For example, within "*Project Area*", some categories have consistently high SHAP values, meaning projects in those areas are frequently associated with rejections, while others lower rejection probability.
- "*Billing Office*" and "*Working Location*" follow a similar trend, suggesting that certain offices or regions handle billing differently, leading to systematic discrepancies in discount patterns.
- Temporal features like "*Work Date Quarter*" and "*Orig Create Date Quarter*" indicate that specific periods in the year correspond to increased rejection likelihood, possibly aligning with budget cycles or financial planning constraints.

Numerical features Such as "*Net Amount*", "*Project Amount*" and "Hours" influence rejection likelihood through magnitude, meaning higher values generally correlate with a greater probability of bill rejection. This trend is visually confirmed in the SHAP plot, where high feature values (red) predominantly push predictions toward higher SHAP values, increasing rejection risk.

Interaction and Polynomial Features: The interaction term "Standard Rate * Project Rate", along with other polynomial features like "*Standard Rate²*" and "*Project Rate²*", captures non-linear relationships that influence rejection probability. While their overall SHAP impact is lower, their presence suggests that the model accounts for threshold effects, where extremely high or low rates lead to disproportionately different outcomes.

Temporal and Financial Influences: The moderate SHAP spread for "*Work Date Quarter*" and "*Orig Create Date Quarter*" suggests that the time when work was performed or invoiced plays

a role in discount decisions. This could be due to seasonal fluctuations in client budgets, contract renewals, or financial review cycles where clients are more inclined to negotiate discounts.

"*Tax*" exhibits moderate horizontal spread, indicating that its overall influence on the model's precision is significant, but limited. While it shows some variance in SHAP contributions, it does not consistently push predictions in one direction. This suggests that tax-related adjustments may play a role in specific billing scenarios but are not as strong or consistent a driver of rejection probability across the dataset as net amounts.

Strategic Implications of SHAP Analysis: The SHAP analysis underscores the dominant role of financial, geographic, and temporal factors in bill rejection decisions.

- "*Project Area*", "*Billing Office*", and "*Working Location*" emerge as critical influencers, reinforcing the importance of regional standardization in billing practices to reduce inconsistencies in discounting.
- "*Net Amount*" and "*Project Amount*" exhibit the strongest financial impact, confirming that large invoices require additional scrutiny and documentation to avoid disputes.
- Temporal trends suggest that certain quarters see more frequent rejections, indicating a need for seasonal billing strategy adjustments.
- Interaction terms and polynomial features contribute moderately to predictions, revealing underlying non-linear effects that shape invoice acceptability.

Narrative Importance: While the narrative content (Description Category) contributes to the model's predictions, it is not among the top-ranked features. The SHAP plot shows that its influence on bill rejection likelihood is limited compared to financial metrics such as "*Net Amount*" and "*Project Amount*", or categorical features like "*Project Area*" and "*Billing Office*". This suggests that while narratives may factor into decision-making, they do not carry the same predictive weight as concrete numerical and categorical attributes, indicating that structured financial and project-related data remain the primary determinants of billing outcomes. However, this may change with even better and more advanced text processing (like state-of-the-art embedding models).

4.4 Analysis of Narrative Impact

The impact of work narratives on bill rejection was a key focus of this research. Over the past couple of years, it has been widely believed by law firms of all sizes that narrative structure, content, length and hygiene are vital to maintaining healthy relationships with clients and avoiding bill rejections and discounts (International Legal Technology Association, 2011). This belief has driven firms to implement strict narrative guidelines and dedicate significant resources to training timekeepers. However, this study sought to evaluate whether narratives truly influence billing outcomes or if other invoice features carry more weight.

Analyzing narrative impact involved using Natural Language Processing (NLP) techniques to categorize and assess work descriptions. TF-IDF (Term Frequency-Inverse Document Frequency) vectorization was ultimately selected as the preferred technique for transforming narrative descriptions into structured categories despite exploring different methods. Although more sophisticated natural language processing models, such as the Sentence Transformer model using '*all-MiniLM-L6-v2*', were evaluated, TF-IDF provided several critical advantages in the context of this research. Most importantly, TF-IDF processing was significantly faster, providing near-instantaneous transformation and clustering of over 160,000 narrative entries, whereas the Sentence Transformer approach required substantially longer computational times and heavier resource consumption.

Given that one of the primary goals of this thesis was to develop a lightweight, efficient tool that could be easily deployed within any law firm environment -regardless of computational infrastructure-, TF-IDF aligned better with the practical demands of real-world implementation. Law firms often need tools that can deliver rapid insights without requiring specialized hardware such as high-end GPUs or dedicated machine learning servers. TF-IDF's speed and efficiency ensure that the narrative categorization step remains accessible to a broader range of firms, from large multinational practices to smaller local offices with limited IT resources.

Furthermore, TF-IDF offers an additional benefit of greater interpretability. When

clustering narratives into categories, the top terms generated through TF-IDF provided a clear and understandable description of each group. This transparency is valuable because it allows billing teams to quickly grasp the kinds of narrative content that tend to be favorably received by clients, facilitating actionable improvements in timekeeper billing practices.

In contrast, while Sentence Transformer embeddings can capture deeper semantic relationships between narratives, they produce abstract vector representations that do not immediately lend themselves to human-readable category descriptions. In the context of legal billing, where stakeholders often need tangible, easily communicated insights rather than complex semantic models, TF-IDF's directness in labeling clusters made it the more practical and business-friendly choice. Ultimately, while newer NLP models offer theoretical performance gains, the marginal improvement in predictive performance observed with Sentence Transformer did not justify the increased complexity and computational cost compared to TF-IDF, especially considering the thesis' objective of creating a usable and efficient predictive tool.

Once transformed using TF-IDF, the narratives were grouped using K-Means clustering, which categorized them into 15 distinct groups based on textual similarities. This method allowed for a more structured comparison of billing outcomes across different types of work descriptions, identifying patterns in how clients respond to specific kinds of legal work.

Clustering Results and Narrative Patterns

The clustering analysis provided new insights into how different categories of narratives correlate with billing acceptance and rejection rates:

- Some categories were consistently linked to bill acceptance. Work descriptions categorized as "Tax Review Regarding" and "Prepare Discuss Email" appeared frequently in approved invoices. These categories contained highly specific, structured, and legally substantive descriptions, reinforcing the idea that clients respond positively to detailed narratives that clearly outline the work performed.
- More generic narratives had a higher likelihood of rejection. Descriptions that lacked

precision, such as "Review Draft" or "Discuss Team Follow-Up," were found in clusters with a higher proportion of rejected invoices. This suggests that vague or repetitive wording may contribute to skepticism from clients, particularly when the value of the task is not clearly conveyed.

- Some categories were neutral or had mixed outcomes. Certain clusters contained descriptions related to administrative or routine legal work, which did not strongly influence billing decisions in either direction.

Figure 9 shows the top groups with their categories and keywords.

```
Category: Review Draft Revise
Keywords: review, draft, revise, documents, comments, report, diligence, advice, client, team

Category: Tax Review Regarding
Keywords: tax, review, regarding, email, discuss, comments, draft, analysis, team, bm

Category: Update Status Review
Keywords: update, status, review, report, documents, draft, email, regarding, send, team

Category: Regarding Review Team
Keywords: regarding, review, team, conference, correspond, draft, analyze, client, bm, follow

Category: Prepare Review Draft
Keywords: prepare, review, draft, regarding, documents, participate, email, team, client, report

Category: Attend Meeting Discuss
Keywords: attend, meeting, discuss, team, prepare, client, conference, regarding, review, bm

Category: Agreement Review Draft
Keywords: agreement, review, draft, comments, revise, regarding, settlement, purchase, email, lease

Category: Letter Draft Review
Keywords: letter, draft, review, regarding, revise, response, prepare, email, client, comments

Category: Email Regarding Review
Keywords: email, regarding, review, draft, send, client, consider, bm, documents, response
```

Figure 9 – Narrative Categories and Keywords

Narrative Influence in the Context of Other Billing Factors

One of the most significant findings of this research is that while narratives do affect billing outcomes, their role is often secondary to financial and structural features of the invoice.

Feature importance analysis revealed that variables such as billing rate, total amount, discount percentage, and project type hold greater predictive power in determining whether a bill

is rejected. While this does not diminish the importance of well-structured narratives, it challenges the assumption that billing success hinges primarily on narrative clarity. Instead, the study suggests that narratives function as a supporting factor rather than a leading determinant.

That being said, narrative precision and specificity still contribute to billing acceptance. Clients appear to be more receptive to invoices that contain detailed, well-organized descriptions of the work performed, rather than vague or redundant language. However, firms may need to reassess how much emphasis they place on narrative refinement compared to optimizing financial structures and billing policies.

Implications for Law Firms

The results suggest that law firms should recalibrate their approach to bill optimization. While efforts to improve narratives remain valuable, they should not come at the cost of neglecting more critical billing factors. Striking a balance between financial accuracy, project classification, and narrative clarity may provide the most effective strategy for minimizing bill rejections.

Ultimately, this study challenges traditional assumptions about the role of narratives in legal billing, indicating that while they do influence client perception, their impact is not as significant as widely believed. Firms that shift their focus toward holistic improvements rather than solely refining narratives may achieve greater success in reducing rejections and maintaining strong client relationships.

4.5 Model Threshold Selection

One of the critical decisions in deploying the predictive model is selecting the appropriate threshold for making predictions.

A systematic approach was employed to determine the optimal threshold. Rather than relying on a single default threshold, a range of thresholds from 0 to 1 was tested, with increments of 0.01. For each threshold, key metrics such as accuracy, precision, and recall were calculated. This process aimed to identify the threshold that provided the best balance, maximizing accuracy

while maintaining an appropriate trade-off between false positives and false negatives. This method allowed for the selection of a threshold that optimizes the model's performance, ensuring it produces the most reliable predictions possible.

- **Optimal Threshold:** The analysis revealed that a threshold of 0.27 provided the best balance between precision and recall for this dataset. At this threshold, the model achieved a high precision of 77.13% while maintaining a recall rate of 95.46%. This threshold strikes a balance between minimizing false positives and ensuring that true rejections are not missed, making it an optimal choice for practical application.
- **Impact of Threshold Selection:** Adjusting the threshold can significantly change the model's predictions. A lower threshold increases recall but decreases precision, meaning the model will catch more true rejections but also produce more false positives. Conversely, a higher threshold increases precision at the cost of recall, leading to fewer false positives but potentially missing some true rejections.

Figure 10 shows the tradeoff between Accuracy, Precision and Recall for different thresholds, as well as the values of the metrics for the selected threshold.

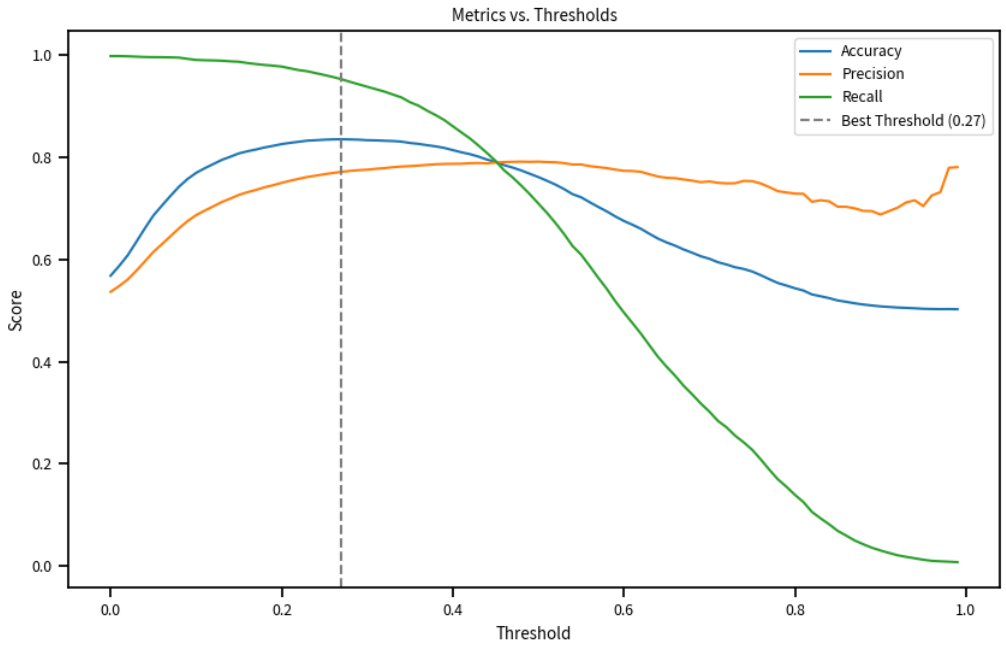

Figure 10 – Metrics vs. Threshold Plot

4.6 Practical Implications of Findings

The findings of this study have several practical implications for law firms, providing actionable insights that can be directly applied to improve billing practices and client relations.

- **Proactive Billing Adjustments:** By understanding which factors most strongly influence bill rejection, law firms can proactively adjust their billing practices to minimize the risk of rejection. For example, ensuring that project areas with historically high rejection rates are carefully reviewed before billing could reduce the likelihood of disputes. Firms can also use these insights to adjust billing rates, structure project teams, and optimize resource allocation.
- **Enhanced Client Communication:** The insights from the SHAP plot and narrative analysis can inform how law firms communicate with clients. For instance, providing clear pricing guidelines or detailed and specific narratives that align with the categories associated with acceptance. These recommendations can improve client satisfaction and reduce the chances of rejection. Firms might develop standardized templates for common tasks or invest in training to ensure that all narratives meet a high standard of clarity and relevance.
- **Risk Management:** The predictive tool developed as part of this thesis offers a practical solution for managing billing risk. By integrating this tool into their billing process, law firms can assess the likelihood of rejection before they send bills to clients, allowing for timely interventions. This proactive approach can help firms avoid costly disputes, improve cash flow, and enhance overall client satisfaction.

The results highlight the significant potential of predictive analytics in transforming legal billing practices. By leveraging the insights gained from this research, law firms can not only reduce the incidence of bill rejections but also foster stronger, more transparent relationships with their clients. The predictive model, coupled with thoughtful implementation, offers a dynamic solution for navigating the complexities of legal billing in today's data-driven environment.

4.7 Tool Development and Implementation

In addition to developing the predictive model, the study entailed creating a practical tool that allows law firms to input new billing data and receive real-time predictions regarding the likelihood of bill rejections. The tool features a user-friendly interface, making it straightforward for non-technical users to upload their datasets, set custom prediction thresholds, and generate output files containing the predicted outcomes.

The tool incorporates the final XGBoost model and provides functionality for:

- **Model Loading:** Users can load the pre-trained model for use with new data. This feature ensures that the tool can be easily updated with improved models as they become available. Figure 11 shows the Model Loading user interface of the tool, where the user must select a XGBoost model in the JSON format. Only the files in the correct format will show and the tool will not proceed unless given the correct file type.

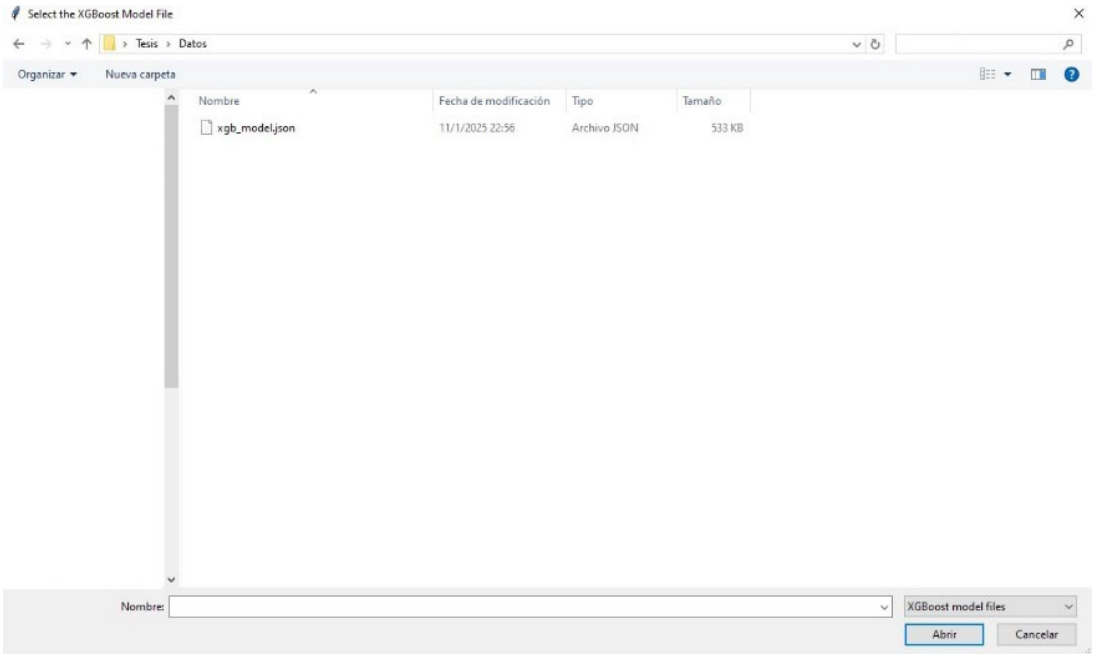

Figure 11 – Prediction Tool: Model Loading

- **Data Input:** The tool accepts datasets in various formats (CSV, text files), preprocessing them automatically to match the format required by the model. This feature simplifies the process for users, allowing them to focus on interpretation rather than data preparation.

Figure 12 shows the Data Input user interface of the tool, where the user must select the appropriate file type to advance.

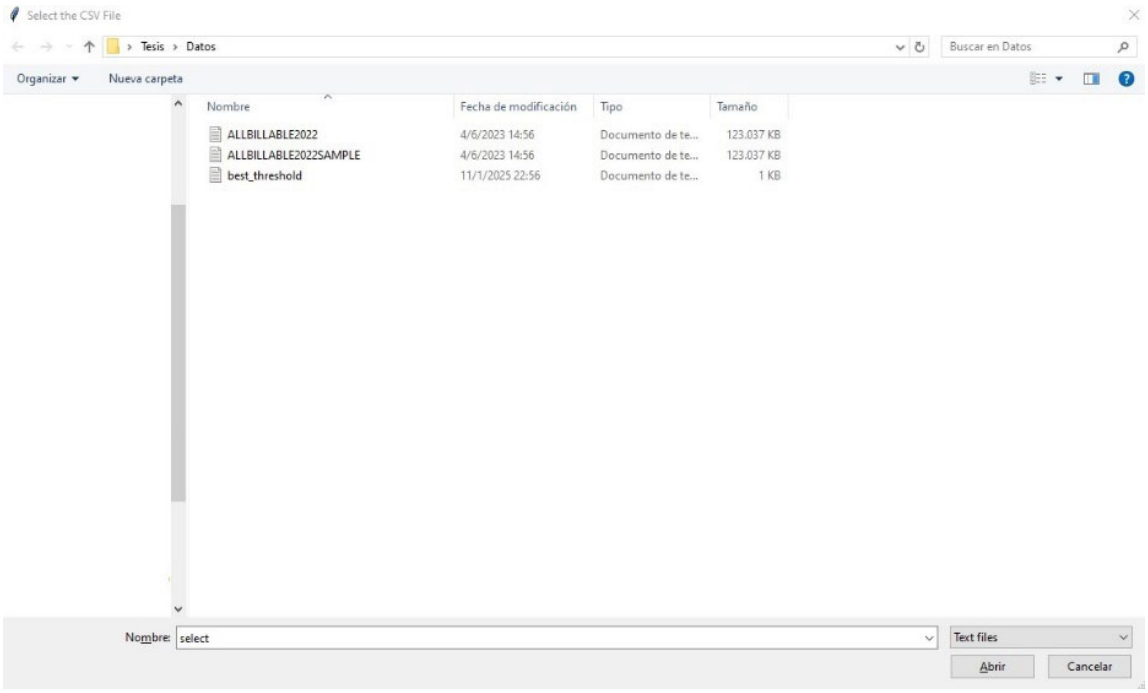

Figure 12 – Prediction Tool: Data Input

- **Threshold Setting:** Users can set a custom threshold to balance the trade-off between precision and recall based on their specific risk tolerance. This flexibility allows law firms to adjust the tool's sensitivity to better match their business needs. Figure 13 shows the Threshold Selection user interface of the tool.

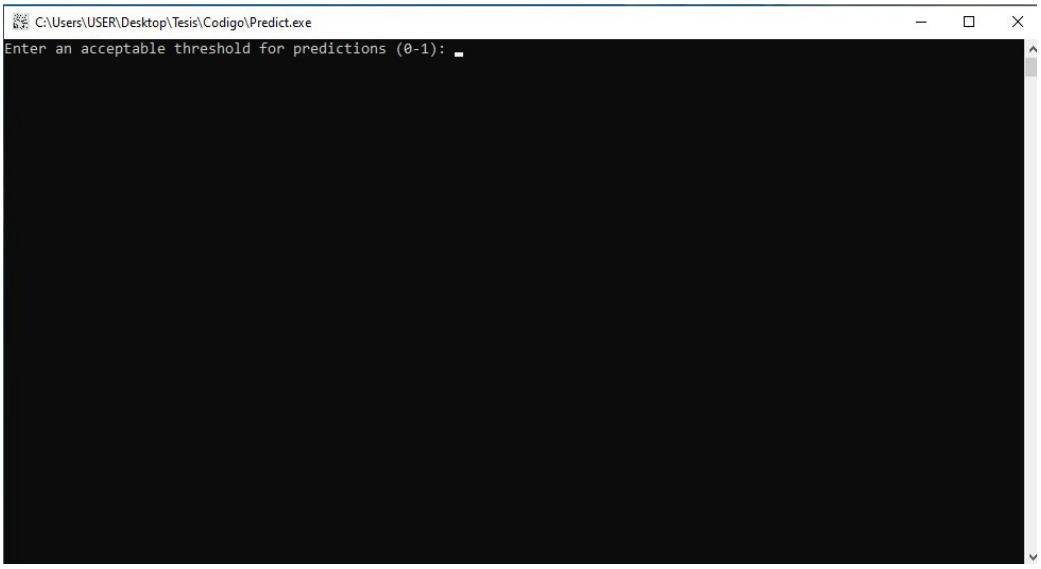

Figure 13 – Prediction Tool: Threshold Selection

- **Prediction Output:** The tool generates a text file containing predictions for each bill in the input dataset, indicating whether the bill is likely to be accepted or discounted. Figure 14 shows the Predictions Exporting user interface of the tool.

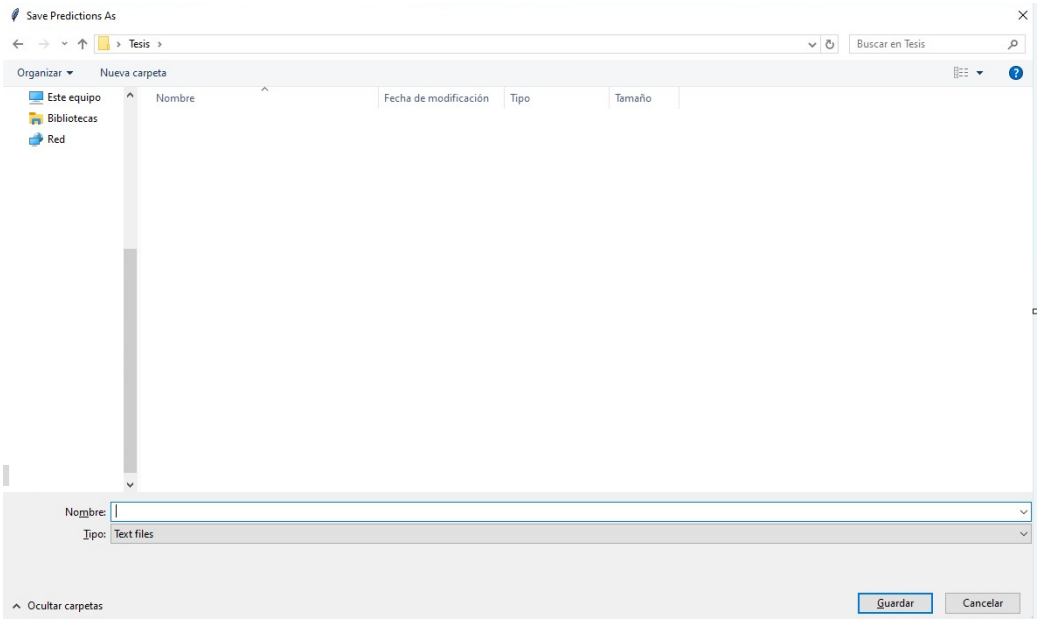

Figure 14 – Prediction Tool: Predictions Exporting

The tool's predictions are exported in a .txt file as a prediction for each row on the dataset. In the file, there will be a row number, followed by a 0 for rows in which the model predicts there will be no rejection and a 1 for rows in which the model predicts a rejection. Figure 15 shows a notepad with the tool's predictions exported in a txt file.

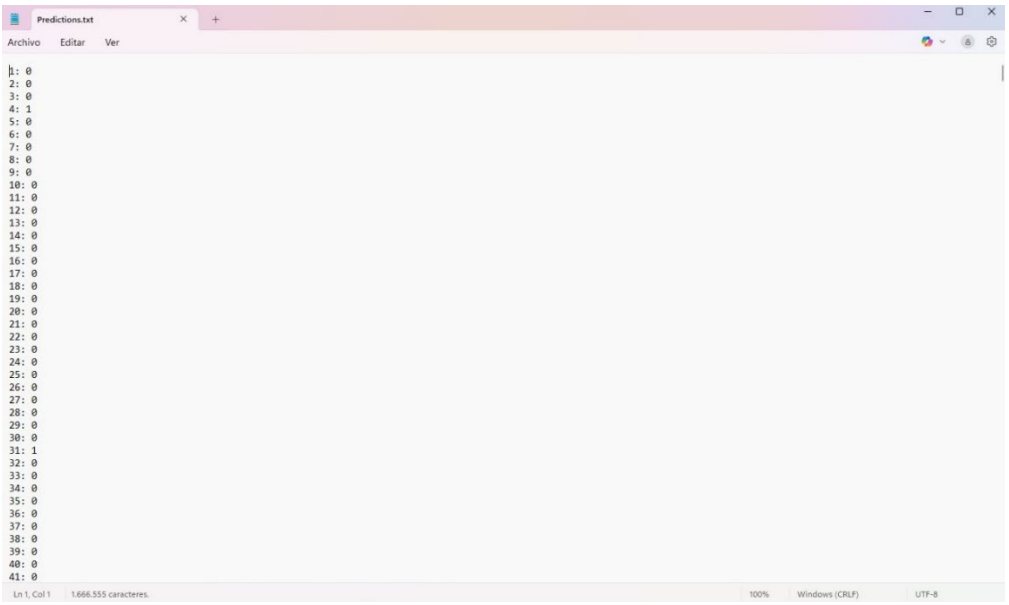

Figure 15 – Prediction Tool: Predictions Output

This tool represents a practical application of the research findings, offering law firms a data-driven approach to optimizing their billing practices and reducing the incidence of bill rejections. A detailed overview of the tool's logic and operational workflow is provided in pseudocode form in Appendix B.

5. Discussion

5.1 Interpretation of Results

The results of this thesis underscore the potential of predictive analytics (specifically machine learning models like XGBoost) in forecasting bill rejections within legal billing processes. The model developed in this study showed remarkable predictive capabilities, with high accuracy, precision, recall, and AUC-ROC scores, which supports its potential utility as a tool for law firms aiming to minimize financial risks associated with bill rejections.

The high accuracy rate of approximately 83.58% suggests that the model can reliably distinguish between bills that are likely to be accepted and those that are likely to be rejected. The model's precision of 77.13% reflects its capacity to minimize false positives, thereby allowing firms to take preemptive actions to address potential issues before submission.

More importantly, the recall rate of 95.46% indicates that the model can identify a substantial proportion of true bill rejections, which is particularly relevant in a legal context where missing a potential rejection could result in financial adjustments or client dissatisfaction. The model's consistent performance across these metrics suggests that it could be integrated into existing billing systems to enhance decision-making processes and improve the accuracy and reliability of billing practices.

One of the most significant insights from the feature importance analysis is the identification of key factors influencing bill outcomes. Contrary to traditional beliefs that emphasize narratives as the primary determinant of bill acceptance or rejection, this study reveals that other factors, such as project and employee area, billing office, and the amounts involved play more critical roles. This finding challenges the conventional wisdom that narrative content alone drives client decisions regarding bill acceptance. Instead, it suggests that law firms should adopt a more comprehensive approach, considering a wider range of operational and financial factors when preparing bills for clients.

It shows that the nature and location of the work, along with the personnel involved, have

a significant impact on the likelihood of bill rejection. This insight could lead to strategic adjustments in how law firms allocate resources, manage projects, operational workflows, and structure their billing practices to minimize the risk of rejection.

Furthermore, the study provides valuable insights into the role of billing offices and employee areas in the billing process. The significant influence of these factors on bill outcomes suggests that regional practices, local client expectations, and the specific expertise of employees in certain areas may play a crucial role in client billing decisions. Law firms could use this information to tailor their billing practices based on regional and personnel-specific factors, ensuring that bills are more likely to align with client expectations and thus be accepted.

5.2 Limitations of the Study

Despite the positive results, this study has several limitations that should be acknowledged:

- **Data Quality and Availability:** The predictive model's accuracy is heavily dependent on the quality and completeness of the data used for training and evaluation. Any inaccuracies or gaps in the billing records could affect the model's predictions. For example, missing data on billing amounts or project details could lead to incorrect predictions.
- **Model Interpretability:** While XGBoost is known for its predictive power, it is also recognized as a "black-box" model, meaning that its internal decision-making processes are not easily interpretable. Although tools like SHAP values provide insights into feature importance and the model's reasoning, the complexity of the model can still pose challenges for stakeholders who require a clear understanding of how decisions are made. This lack of transparency might hinder the adoption of the model in firms where interpretability is crucial for decision-making.
- **Narrative Analysis:** The approach to narrative analysis in this study involved clustering and categorizing narratives using relatively straightforward techniques like TF-IDF

vectorization and K-Means clustering. While effective for the purposes of this research, these methods may oversimplify the nuances of legal work descriptions. Legal narratives are often complex and context- dependent, and the current approach may not fully capture the subtleties involved. More sophisticated natural language processing (NLP) techniques could potentially yield more accurate and nuanced insights.

- **Scope of Application:** The model's application is currently limited to predicting bill rejections based on historical data from a specific legal context. Though the dataset used for the analysis spans multiple areas of law and different practices, the applicability of the model to other areas or different types of legal work remains to be explored.
- **Generalizability:** Moreover, the focus on a single law firm's data means that the model may not account for variations in billing practices across different jurisdictions or areas of law. Law firms with unique practices or clientele might require further customization of the model or even retraining on their specific data to achieve similar predictive accuracy.
- **Technological Integration:** Implementing the model within existing legal billing systems might pose challenges, especially for firms with legacy systems or those lacking the necessary technological infrastructure. This could limit the model's adoption and effectiveness, particularly in smaller firms with limited resources for technological upgrades. Furthermore, legal professions are in the top two profession groups with the highest exposure to being replaced by AI and automation in the United States (Briggs et al., 2023), and the industry perceives itself as not having the best scores in data collection and technology adoption. They also see organizational barriers preventing them from investing in technology, expecting to make greater use of it to improve productivity sooner rather than later (Wolters Kluwer, 2023).
- **Applicability to Small and Medium-Sized Law Firms:** Another important limitation is the potential lack of direct applicability of these findings to small or medium-sized law firms. The dataset analyzed in this research was sourced from a large, international law firm with

complex billing structures, a high volume of transactions, and formalized internal processes. In contrast, smaller firms often rely on simpler billing systems, fewer staff, and closer, more informal relationships with their clients. These differences in organizational structure, billing practices, and client dynamics may affect both the factors that lead to bill rejections and the feasibility of implementing advanced predictive analytics tools. As a result, the model developed, and the insights generated in this thesis may require adaptation or simplification before they can provide meaningful value in smaller practice settings. Further research involving a broader range of firm sizes and operational models may be necessary to ensure that predictive analytics solutions can be effectively tailored to the unique needs and constraints of small and medium-sized law firms.

5.3 Suggestions for Future Research

To build on its findings and address its limitations, this thesis proposes several avenues for future research:

- **Broader Data Collection:** Future studies should aim to expand the dataset to include billing records from multiple law firms across different jurisdictions and areas of practice. This would improve the generalizability of the model and provide a more comprehensive understanding of the factors influencing bill rejections across the legal industry. Additionally, incorporating data from firms with diverse client bases could help in identifying patterns specific to certain types of clients or legal work.
- **Advanced Natural Language Processing (NLP):** The narrative analysis could be significantly enhanced by incorporating more advanced NLP techniques, such as deep learning-based language models like BERT (Devlin et al., 2019) or GPT (Radford et al., 2018). These models have shown great promise in understanding complex text and could provide deeper insights into the role of work descriptions in billing decisions. By capturing the context and subtleties of legal narratives, these techniques could improve the model's ability to predict bill rejections based on narrative content.

- **Model Interpretability:** Future research could explore methods to improve the interpretability of complex models like XGBoost. Techniques such as LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016) or the development of simpler, more transparent models could be considered. These approaches would make the model's predictions more accessible to stakeholders who require a clear understanding of the decision-making process.
- **Integration with Other Predictive Tools:** There is potential for integrating the bill rejection prediction model with other predictive tools used in legal practice, such as client retention models or litigation outcome predictors[5]. Such integration could lead to the development of a comprehensive suite of predictive analytics solutions for law firms, providing a holistic view of their operations and client interactions. This could enhance decision-making across various aspects of legal practice, from billing to case management.
- **Exploration of Ethical Implications:** As predictive analytics become more prevalent in legal practice, it is important to consider the ethical implications of these technologies. Future research should explore issues such as data privacy, algorithmic bias, and the potential for misuse of predictive models. Ensuring that predictive tools are used responsibly and ethically is crucial for maintaining public trust, safeguarding client confidentiality, and upholding the integrity of legal outcomes. One significant concern is the potential for these algorithms to reinforce existing biases in historical billing data. If certain types of legal work, practice areas, or geographic locations have historically experienced higher rejection rates, the model may learn to flag them as higher risk, unintentionally discouraging firms from taking on these matters. Similarly, clients with frequent past rejections might be deprioritized or avoided altogether, leading to systemic exclusion. Without proper oversight, these predictive tools could create an unfair disadvantage for specific groups, improving efficiency but preventing equal treatment.

[5] Litigation prediction software such as Lex Machina, Premonition AI, and CaseMine.

6. Conclusion

This thesis explored the application of predictive analytics to the specific domain of legal billing, with a particular focus on forecasting bill rejections. The development and evaluation of a predictive model using XGBoost have provided new insights into the factors that most significantly influence bill acceptance or rejection. This research contributes to a deeper understanding of the operational dynamics within law firms and offers a practical tool that can be implemented to enhance billing practices, reduce financial risk, and improve client relationships.

One of the most impactful findings of this research is the identification of key features that influence bill rejections. Contrary to traditional beliefs that emphasize the importance of work narratives, this study found that other factors, such as project area, billing office, and employee roles, play a more substantial role in determining whether a bill is accepted or rejected. This insight encourages law firms to broaden their focus beyond just narrative quality and to consider a wider range of operational factors when preparing bills. The ability to identify and address these critical factors before bills are submitted can lead to more accurate billing practices, thereby enhancing client satisfaction and reducing the likelihood of costly rejections.

6.1 Strategic Implications for Law Firms

The predictive tool developed through this research represents a strategic asset for law firms. By integrating this tool into their billing processes, firms can achieve greater accuracy and efficiency in their operations. The ability to predict bill rejections not only reduces the time and resources spent on resolving disputes but also strengthens the firm's financial management. This is particularly important in a competitive legal market where operational efficiency and client satisfaction are key differentiators.

Moreover, the insights gained from the feature importance analysis offer law firms a data-driven basis for refining their billing strategies. For instance, understanding that certain project areas or billing offices are more prone to rejections allows firms to implement targeted interventions, such as additional reviews or consultations with clients before bills are finalized.

This level of strategic adjustment, informed by predictive analytics, can lead to a more robust and resilient billing process that aligns more closely with client expectations and reduces the risk of rejection.

6.2 Broader Impacts on the Legal Industry

Beyond the immediate implications for legal billing, this thesis illustrates the broader potential of predictive analytics in the legal industry. The success of the XGBoost model in accurately forecasting bill rejections underscores the value of machine learning in analyzing complex legal data. This approach can be extended to other areas of legal practice, such as predicting case outcomes, managing contract risks, and optimizing resource allocation. The ability to leverage historical data to inform future decisions marks a significant shift in how legal services can be delivered more effectively and efficiently.

The introduction of predictive analytics into legal billing also encourages a cultural shift within law firms toward data-driven decision-making. As firms become more accustomed to using predictive tools, they may begin to explore other areas where data analytics can drive improvements. This could lead to a more integrated approach to legal operations, where insights from different predictive models are used in concert to enhance overall firm performance. Such a shift would position law firms to better meet the demands of modern clients, who increasingly expect transparency, efficiency, and value from their legal service providers.

6.3 Enhancing Client Relationships

One of the most significant impacts of implementing the predictive model is its potential to enhance client relationships. Billing disputes are a common source of tension between law firms and their clients. By reducing the likelihood of rejections through more accurate billing practices, law firms can foster greater trust and satisfaction among their clients. The predictive model allows firms to identify and address potential issues before they escalate into disputes, reflecting a commitment to delivering value and transparency.

Additionally, the insights from narrative analysis, while not the most critical factor in this

study, still play a role in client communication. The ability to categorize and optimize narratives according to client preferences can further strengthen client relationships. Firms that take the time to understand and align with client expectations, as informed by the predictive model, are likely to see improved client retention and satisfaction. This not only benefits the firm's reputation but also contributes to long-term financial stability.

6.4 Future Directions for Practice and Innovation

Beyond expanding the academic understanding of bill rejection dynamics, future efforts should focus on translating predictive insights into operational tools that enhance decision-making in legal practice. One promising direction involves embedding the predictive model into existing billing software or legal management platforms, enabling real-time risk assessments during invoice preparation. This integration would allow legal professionals to receive proactive recommendations on how to improve billing narratives, adjust fee structures, or flag high-risk entries before submission.

Additionally, law firms could leverage predictive outputs to inform internal process improvements, such as refining billing review workflows, identifying training needs for specific teams, or establishing escalation protocols for at-risk invoices. Over time, historical feedback from accepted and rejected bills could be used to retrain and personalize models, allowing firms to develop firm-specific predictors aligned with their clients' expectations and practices.

Another area for development is the incorporation of visual analytics dashboards, enabling lawyers and finance teams to explore rejection risks interactively. These tools could provide actionable insights at both the individual invoice level and the portfolio level, supporting more strategic conversations around pricing, staffing, and resource allocation.

6.5 Conclusion

In conclusion, the findings of this thesis illustrate the practical value of predictive analytics in legal billing. The development of a robust predictive model using XGBoost can provide law firms with a powerful tool to forecast bill rejections, optimize billing practices, and enhance client

satisfaction. As the legal industry continues to evolve, the adoption of data-driven approaches will be critical to staying competitive and meeting the growing demands of clients. This research contributes to that evolution, offering both a practical solution and a foundation for future innovation in legal practice. It also shows that, even though how legal firms present their work to their clients through narratives is important, other aspects of legal billing carry more weight when it comes to rejections and discounts.

7. References

7.1 Books and Academic Journals

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Coenen, L., Verbeke, W., & Guns, T. (2022). *Machine learning methods for short- term probability of default: A comparison of classification, regression, and ranking methods*. Journal of the Operational Research Society, *73*(1), 191-206. https://doi.org/10.1080/01605682.2020.1865847

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly Media.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York, NY, USA: Springer. https://doi.org/10.1007/978-0-387-84858-7

Kowsari, K., Heidarysafa, M., Brown, D. E., Meimandi, K. J., & Barnes, L. E. (2019). *Text classification algorithms: A survey.* Information, 10(4), 150. https://doi.org/10.3390/info10040150

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge, UK: Cambridge University Press.

Molnar, C. (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub. Retrieved April 30, 2025, from https://christophm.github.io/interpretable-ml-book/

Powers, D. M. W. (2011). *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation.* Journal of Machine Learning Technologies,

2(1), 37-63. Retrieved from https://arxiv.org/pdf/2010.16061

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. (2017). *Data mining for business analytics: Concepts, techniques, and applications with XLMiner.* Hoboken, NJ, USA: Wiley.

Susskind, R. (2017). *Tomorrow's lawyers: An introduction to your future* (2nd ed.). Oxford, UK: Oxford University Press.

7.2 Conference Papers

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). San Francisco, CA: ACM. https://doi.org/10.1145/2939672.2939785

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186). Minneapolis, MN, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. Proceedings of the International Conference on Learning Representations (ICLR). Retrieved April 30, 2025, from https://arxiv.org/abs/1301.3781

Ramos, J. (2003). *Using TF-IDF to determine word relevance in document queries.* Proceedings of the First Instructional Conference on Machine Learning (pp. 133-142). Piscataway, NJ, USA: Department of Computer Science, Rutgers University.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). San Francisco, CA, USA: ACM. https://doi.org/10.1145/2939672.2939778

7.3 Industry Reports and White Papers

Briggs, J., Kodnani, D., Hatzius, J., & Pierdomenico, G. (2023). *The potentially large effects of artificial intelligence on economic growth.* New York, NY, USA: Goldman Sachs Global Economics Analyst. Retrieved April 30, 2025, from https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html

Clio. (2020). 2020 Legal Trends Report. Vancouver, Canada: Clio. Retrieved April 30, 2025, from https://www.clio.com/resources/legal-trends/2020-report/

International Legal Technology Association (ILTA). (2011). *Financial management: A slice of the finance pie.* Austin, TX, USA: ILTA White Paper. Retrieved April 30, 2025, from https://epubs.iltanet.org/i/30285-financial-management/0?

Matich, T. (2019). *Law Firm Billing: Ultimate Guide and Best Practices*. Retrieved April 30, 2025, from Clio: https://www.clio.com/blog/law-firm-billing/

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training.* San Francisco, CA, USA: OpenAI. Retrieved April 30, 2025, from https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Thomson Reuters. (2020). *How AI and machine learning is shaping legal strategy*. Retrieved April 30, 2025, from Thomson Reuters Careers Blog: https://www.thomsonreuters.com/en/careers/careers-blog/how-ai-and-machine-learning-is-shaping-legal-strategy.html

Thomson Reuters. (2023). *Law firm billing efficiency and write downs: A Thomson Reuters study.* Toronto, Canada: Legal Executive Institute. Retrieved April 30, 2025, from https://www.thomsonreuters.com/en-us/posts/wp-content/uploads/sites/20/2024/04/Law-Firm-Billing-Write-Downs-2023-1.pdf

Wolters Kluwer. (2023). *Future Ready Lawyer Survey Report 2023: Embracing Innovation, Adapting*

to Change. Alphen aan den Rijn, Netherlands: Wolters Kluwer. Retrieved April 30, 2025, from https://www.wolterskluwer.com/en/know/future-ready-lawyer-2023

Yu, J., Elmankabady, K., Liu, Z., & Kwong, A. (2022). *Predictive analytics.* Toronto, Canada: Future of Law Lab, University of Toronto Faculty of Law. Retrieved April 30, 2025, from https://futureoflaw.utoronto.ca/sites/default/files/Website/Predictive%20Analytics%20report%202022.Pdf

7.4 Online Resources and Software Documentation

CaseMine by Gauge Data Solutions Pvt. Retrieved April 30, 2025, from https://www.casemine.com/

KMeans Clustering Documentation. (n.d.). Retrieved April 30, 2025, from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Lex Machina by Lexis Nexis. Retrieved April 30, 2025, from https://lexmachina.com/how-it-works/

Matplotlib Documentation. (n.d.). Retrieved April 30, 2025, from https://matplotlib.org/stable/contents.html

NumPy Documentation. (n.d.). Retrieved April 30, 2025, from https://numpy.org/doc/stable/

Optuna Documentation. (n.d.). Retrieved April 30, 2025, from https://optuna.org/

Pandas Documentation. (n.d.). Retrieved April 30, 2025, from https://pandas.pydata.org/

Premonition AI by Premonition. Retrieved April 30, 2025, from https://premonition.ai/

Python Software Foundation. (2023). Python Language Reference, version 3.9. Retrieved April 30, 2025, from https://www.python.org/

Scikit-learn Documentation. (n.d.). Retrieved April 30, 2025, from https://scikit-learn.org/stable/

SciPy Documentation. (n.d.). Retrieved April 30, 2025, from

https://docs.scipy.org/doc/scipy/

Seaborn Documentation. (n.d.). Retrieved April 30, 2025, from

https://seaborn.pydata.org/

Sentence-Transformers Documentation. (n.d.). Retrieved April 30, 2025, from

https://www.sbert.net/

Sentence-Transformers: all-MiniLM-L6-v2 Model. (n.d.). Retrieved April 30, 2025, from

https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

SHAP Documentation. (n.d.). Retrieved April 30, 2025, from https://shap.readthedocs.io/

XGBoost Documentation. (n.d.). Retrieved April 30, 2025, from https://xgboost.readthedocs.io/

Appendix A. Dataset Variables and Data types

Variable	Data Type
Client_Category	Categorical
Client_Number	Categorical
Client	Categorical
Element	Categorical
Element_Description	Categorical
Project_Number	Categorical
Project_Description	Categorical
Project_Reporting_Group	Categorical
Project_Type	Categorical
Billing_Representative	Categorical
Management_Representative	Categorical
Project_Area	Categorical
Employee_Number	Categorical
Employee_Name	Categorical
Employee_Position	Categorical
Employee_Group	Categorical
Employee_Area	Categorical
Orig_Create_Date	Date/Time
Work_Date	Date/Time
Working_Location	Categorical
Hours	Numerical
Hours_After_Write_Down	Numerical
Standard_Rate	Numerical
Standard_Amount	Numerical
Project_Rate	Numerical
Project_Amount	Numerical
Net_Amount	Numerical

Quantity_Write_Down_Amount	Numerical
Value_Write_Down_Amount	Numerical
Currency	Categorical
Billing_Office	Categorical
Invoice_Number	Textual
USERID	Categorical
Invoice_Item_#	Textual
Billing_Code	Categorical
Billing_Date	Date/Time
Amount_Billed	Numerical
Billing_Grand_Value	Numerical
Billing_Grand_Value_After_Write_Up_Down	Numerical
Billing_Discount_Amount	Numerical
Billing_Fee_Markup	Numerical
Billing_Vat_Surcharge_On_IOB	Numerical
Tax	Numerical
Billing_Currency	Categorical
Cost_Code	Categorical
Cost_Code_Description	Categorical
Task_Code	Categorical
Task_Code_Description	Categorical
Activity_Code	Categorical
Activity_Code_Description	Categorical
FF_Activity_Code	Categorical
FF_Task_Code	Categorical
Cost_Type	Categorical
REF_ID	Categorical
CO_Document_Number	Categorical
CO_Doc_Item_Number	Categorical
FI_Doc_Number	Categorical

Draft_Bill_Number	Categorical
DrBl_Item_Number	Categorical
DB_Rejection	Categorical
RFR_Description	Categorical
Time_Cost_Ind	Categorical
Bill_Status	Categorical
DTE_Identifier	Textual
Has_Consolidation	Categorical
Work_Description	Textual

Appendix Notes:

- Variable names have been modified to preserve confidentiality of original data.

Appendix B. Bill Rejection Prediction Tool Pseudocode

```
BEGIN BillRejectionPredictionTool

 // 1. Model Loading
 PROMPT user to select the pre-trained predictive model file (XGBoost JSON)
 IF model file selected THEN
    LOAD the trained model from file
 ELSE
    PROMPT again until valid input received

 // 2. Data Input
 PROMPT user to select a billing data file (CSV or TXT)
 IF data file selected THEN
    READ the file into memory as a dataset
 ELSE
    PROMPT again until valid input received

 // 3. Data Preprocessing
 FOR each column in the dataset DO
    IF column is categorical THEN
       CONVERT column type to category
    END IF
    IF column is numeric THEN
       CONVERT column to numeric format, handling missing/invalid values
    END IF
 END FOR
 FILTER out rows and drop columns as done in the model training phase
 ENGINEER additional features (e.g., polynomial combinations of numeric fields)
 CONVERT date fields to categorical representations (such as quarters)
 HANDLE missing narrative text entries
 APPLY text vectorization and cluster narratives with unsupervised clustering
 MAP each narrative cluster to a descriptive label
 REMOVE intermediate or unnecessary columns

 // 4. Threshold Selection
 PROMPT user to enter a prediction threshold (value between 0 and 1)
 IF user input is valid THEN
    SET threshold to user-specified value
 ELSE
    PROMPT again until valid input received

 // 5. Prediction Generation
 FOR each row in the preprocessed dataset DO
    COMPUTE probability of bill rejection using the loaded model
    IF probability >= threshold THEN
       SET prediction to 1 (likely rejection)
    ELSE
       SET prediction to 0 (likely acceptance)
    END IF
    RECORD (row index, prediction) in result list
```

```
  END FOR

  // 6. Output Results
  PROMPT user to select a location to save the prediction results
  IF location selected THEN
     WRITE predictions to text file in format: [row_number]: [prediction]
        (1 = likely rejection, 0 = likely acceptance)
     DISPLAY success message
  ELSE
     PROMPT again until valid input received

END BillRejectionPredictionTool
```