Predictive Analytics for Healthcare Claims Processing

Mohamed Alderei & Raunav Sharma

Supervisor: Alexander Pelaez, Ph.D. 5e Analytics

DSCI-D 499 - Muazzam Siddiqui

April 26, 2024

Abstract

This final report encapsulates the comprehensive execution and outcomes of our science capstone project. The project's goal was to enhance financial predictability within healthcare billing systems by applying sophisticated data analytics techniques such as machine learning algorithms and predictive modeling. This endeavor addressed critical challenges encountered by a behavioral healthcare company, focusing on optimizing revenue cycle management. The document thoroughly outlines the project's objectives, methodologies, results, evaluations, and strategic recommendations for future advancements.

Chapter 1: Introduction

1.1 Project Overview

This capstone project was designed to advance financial predictability in healthcare billing systems through sophisticated data analytics methodologies. In collaboration with a behavioral healthcare company, the initiative focused on enhancing revenue forecasting accuracy, developing predictive models for payment timelines, and devising algorithms to identify potential claim denials.

1.2 Background and Importance

Healthcare organizations frequently encounter challenges related to unpredictable payment collections, adversely affecting cash flow and operational efficiency. Addressing these issues is vital for maintaining financial stability and optimizing resource distribution. Utilizing advanced data analytics, this project provided insights into the intricacies of revenue cycles and claims processing, thereby facilitating significant improvements in financial predictability and resource management efficiency.

1.3 Introduction to the Healthcare Domain

The healthcare sector is characterized by intricate processes and terminologies pivotal to revenue cycle management (RCM), billing, and claims processing. Key concepts essential for this project include Current Procedural Terminology (CPT) codes and the aging of claims. Understanding these elements is crucial for developing sophisticated data analytics solutions tailored for healthcare billing.

1.4 Assessment of the Current Solution

Healthcare billing systems currently struggle with uncertainties around the acceptance or denial of claims, largely due to reliance on manual processes and outdated systems. These uncertainties contribute to inefficiencies and inaccuracies within revenue forecasting and claims management. Adopting advanced data analytics can significantly reduce these uncertainties by predicting claim outcomes with greater accuracy, thus optimizing the financial predictability and operational efficiency of healthcare organizations.

1.5 Research Questions

The research questions guiding this project focused on leveraging data analytics to enhance financial predictability in healthcare billing. Key questions included:

1. How can we develop predictive models to accurately forecast future revenue based on historical and current billing data?
2. What predictive models can be implemented to determine the likelihood of claim acceptance or denial, and how can these models improve the predictability and efficiency of healthcare billing systems?

1.6 Business Objectives

The project aimed to develop and deploy sophisticated predictive analytics models to improve financial management within healthcare billing. Specifically, it focused on creating two distinct sets of models: one set for accurately forecasting future revenue and another for predicting the acceptance or denial of claims, thereby optimizing operational efficiencies and resource allocation.

1.7 Business Success Criteria

The success of this project will be determined by the effectiveness of the selected models in achieving the highest precision and accuracy for their designated tasks. Success criteria include achieving high accuracy and low error rates in revenue forecasting to ensure financial stability and effective prediction of claim acceptance or denial to minimize incorrect processing and enhance operational efficiency. Additionally, the adaptability of these models to changes in data patterns and their scalability within the healthcare organization’s billing systems will be crucial.

1.8 Data Mining Objectives

The data mining objectives were focused on evaluating different predictive models to determine the most effective approach for two key tasks: accurately forecasting future revenue and predicting the acceptance or denial of claims. The goal was to select the best model for each task based on performance metrics, thereby enhancing the financial predictability and efficiency of healthcare billing operations.

1.9 Data Mining Success Criteria

For the claim predictability models, success will be evaluated through a comparative analysis where multiple models are assessed based on their accuracy and F1-score. The selection of the best model will be determined by its superior performance on these metrics, supported by preliminary tests such as ROC curves, confusion matrices, and cross-validation. These tests provide initial indicators of each model's effectiveness and robustness but will be explored in depth later in the report to substantiate the final model choice. This methodical approach ensures that the chosen model excels in both technical performance and practical applicability for predicting claim outcomes. Success for the revenue prediction model will be evaluated using the Akaike Information Criterion (AIC) and the Dickey-Fuller test. The AIC will be instrumental in selecting the best model by comparing their fit, with lower values indicating a more favorable model. The Dickey-Fuller test ensures the stationarity of the time series data, a necessary condition for the reliability of predictive models. Together, these tests confirm the chosen model’s suitability for accurately forecasting future revenues in healthcare billing.

1.10 Scope and Limitations

This project is focused on utilizing billing and claims data from a behavioral healthcare company to develop predictive models for forecasting revenue and determining claim acceptance or denial. The scope of this analysis is constrained by several factors. Key limitations include a restricted volume of patient data available due to stringent security and privacy requirements, which impacts the comprehensiveness and depth of the analysis. Additionally, the limited diversity of data points restricts the ability to generalize the findings across different healthcare settings or broader populations.

1.11 Project Plan

The project progressed through three main phases:

1. **Problem Formulation**: Initially, we defined the project's challenges and goals in consultation with the project supervisor, focusing on key issues suitable for predictive analytics, such as revenue forecasting and claim predictability.

2. **Model Development and Implementation**: This phase involved data collection, cleaning, and preparation to optimize the limited data. Various predictive models were developed, tested, and iteratively refined to address the defined challenges effectively.

3. **Model Evaluation and Refinement**: The final stage included rigorous evaluation of each model using specific metrics like AIC and ROC curves. The

best-performing models were fine-tuned through cross-validation and other methods to ensure their accuracy and reliability in operational settings.

Chapter 2: Data

2.1 Data Source

The data for this project was provided by our project supervisor, affiliated with 5e Analytics, a data consulting firm. The dataset originated from a behavioral healthcare company, although specific details about the company were not disclosed to ensure confidentiality. The acquisition of data faced delays, primarily due to the lengthy processes of data anonymization and compliance checks necessary to adhere to privacy regulations and ethical standards. This stage required careful coordination to ensure that the data, once received, maintained its integrity and relevance for the analytical goals of the project.

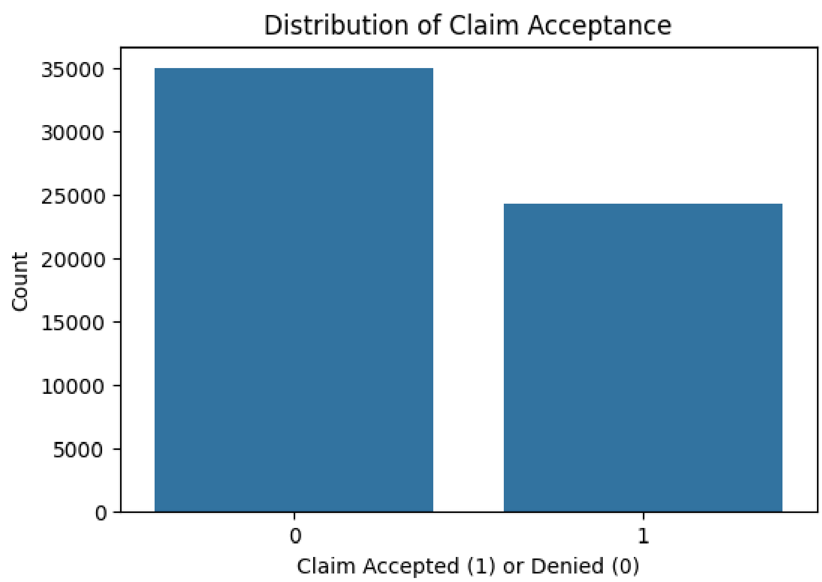
2.2 Data Description

The dataset comprises 213,679 rows and 37 columns, representing a substantial volume of data capable of supporting in-depth analytical insights. Each row in the dataset corresponds to a unique record within the claims processing system, providing a comprehensive snapshot of individual claims. While the dataset includes several columns, the analysis primarily focused on the following key attributes due to their direct relevance to the project's objectives:

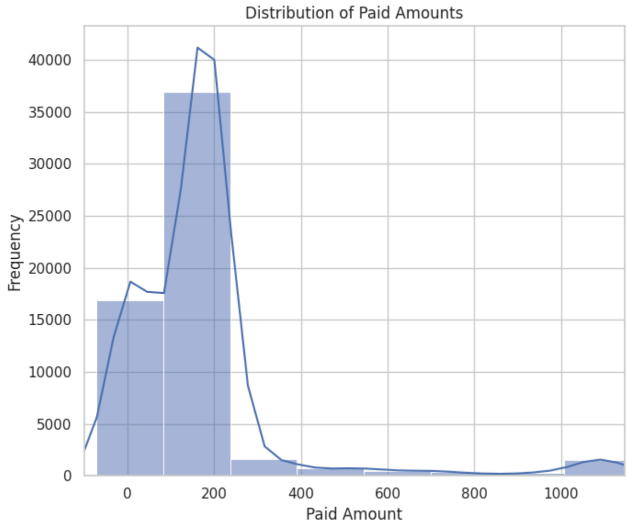
- **Paid Amount:** Represents the monetary value associated with each claim. This attribute is fundamental for conducting revenue analysis, allowing for assessments of financial outcomes related to each claim.
- **Claim Amount:** The total billed amount for each claim, essential for evaluating the financial demands placed by the healthcare provider before adjustments and payments.
- **Explain Code_1:** Indicates the status of each claim, essential for distinguishing between accepted and rejected claims. This attribute plays a pivotal role in analyzing the factors influencing claim adjudication outcomes.
- **Alt_CPT_Code:** Contains alternative codes for procedures related to different healthcare services. This attribute helps in exploring the associations between specific treatments and claim outcomes.
- **Date Billed:** The date on which the claim was initially billed, critical for understanding the billing timelines and analyzing delays in the claims processing lifecycle.
- **Adjudication Date:** The date the claim was adjudicated, providing insights into the duration between billing and resolution, which is crucial for analyzing payment timelines and identifying potential bottlenecks in the claims processing workflow.analysis of different data attributes.

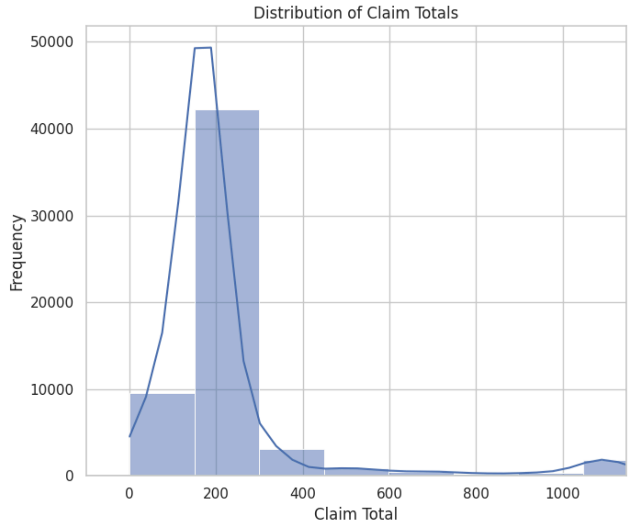
2.3 Data Exploration

In the exploratory data analysis, we delved into various aspects of the dataset to uncover initial patterns and insights:



Distribution of Claims Status: The distribution of claim acceptance showed a notable difference between the number of claims accepted and those denied. From the bar chart, it was observed that the number of accepted claims is substantially higher than the number of denied claims, indicating a trend that may influence the development of predictive models.





Financial Distributions: The histograms for both 'Paid Amount' and 'Claim Amount' demonstrated a right-skewed distribution with a peak for lower amounts, suggesting that most claims involve lesser financial figures. The long tail for both distributions indicates infrequent occurrences of very high claim values.

Statistical Summary: The mean 'Paid Amount' of $205.90 was notably lower than the mean 'Claim Amount' of $245.15. The data exhibited a wide spread as evidenced by the standard deviation values, with $451.35 for 'Paid Amount' and $441.06 for 'Claim Amount'. The presence of negative values in 'Paid Amount' could indicate refunds or claim adjustments. Both distributions showed their maximum at $7481.70, possibly highlighting exceptional cases or outliers in the data

.

These financial insights are critical for understanding the underlying payment dynamics and will be instrumental in shaping the predictive modeling, particularly in feature engineering and in setting the context for the model's expected performance.

2.4 Data Preparation

To prepare the dataset for modeling, we undertook several key steps:

Data Cleaning: We addressed any missing or inconsistent data points found during the quality assessment. For numerical fields like 'Paid Amount' and 'Claim Amount', we used mean or median imputation where appropriate, considering the distribution of the data.

Transformation and Normalization: We normalized the financial amounts to ensure that the model inputs had a standard scale. This helped to improve the convergence time of the models and the reliability of the results.

Feature Engineering: From the initial exploratory analysis, we engineered new features that could enhance model performance. This included indicators for common CPT codes, flags for outlier claims, and calculated fields like claim duration (time between 'Date Billed' and 'Adjudication Date').

Encoding Categorical Variables: Categorical variables such as CPT codes were encoded using one-hot encoding or label encoding to convert them into a format that could be processed by the predictive models.

Data Splitting: The dataset was split into training and testing sets to validate the performance of the models on unseen data, which is critical for assessing their generalization capability.

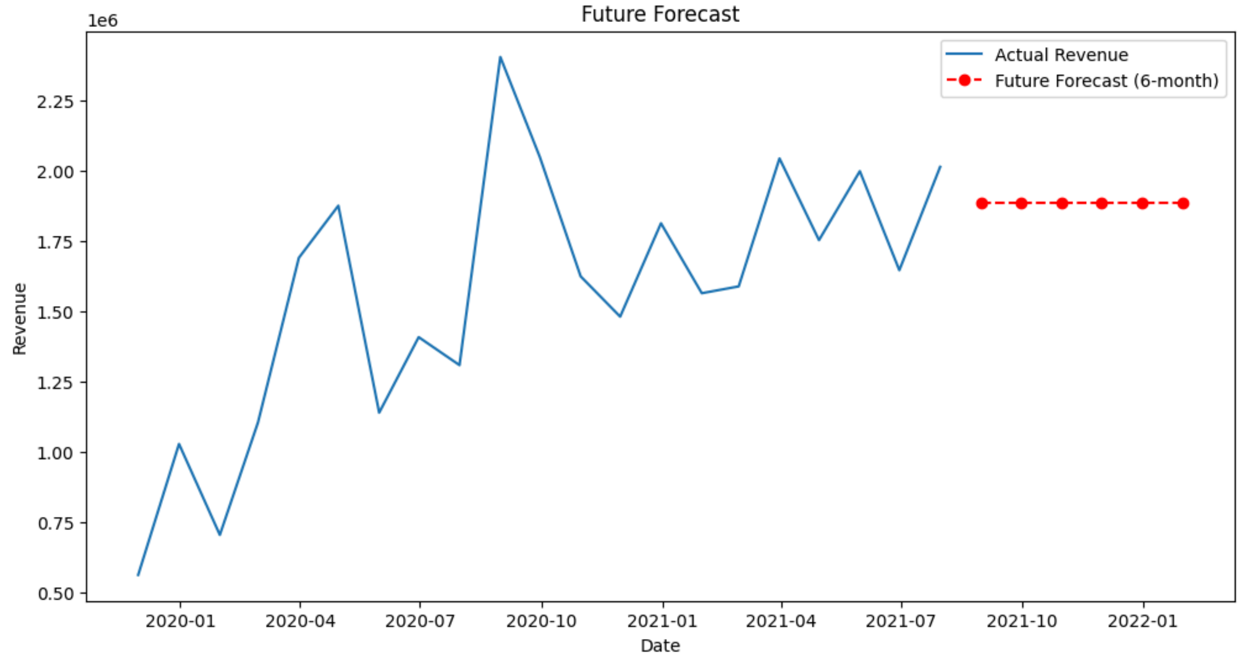
These preparation steps were designed to maximize the quality and utility of the data fed into the modeling algorithms, thereby enhancing the potential for accurate and robust predictive performance.

Chapter 3: Model

In our comprehensive modeling endeavor, we aim to tackle two critical aspects of our business operations: accurate sales forecasting and effective claim prediction. By leveraging a range of analytical techniques, our objective is to unearth hidden patterns within our data, facilitating informed decision-making and robust strategic planning. This report details our journey through developing methodologies that not only predict future sales with precision but also determine the likelihood of claim acceptance or denial, ensuring operational efficiency and financial accuracy.

3.1 Baseline Models

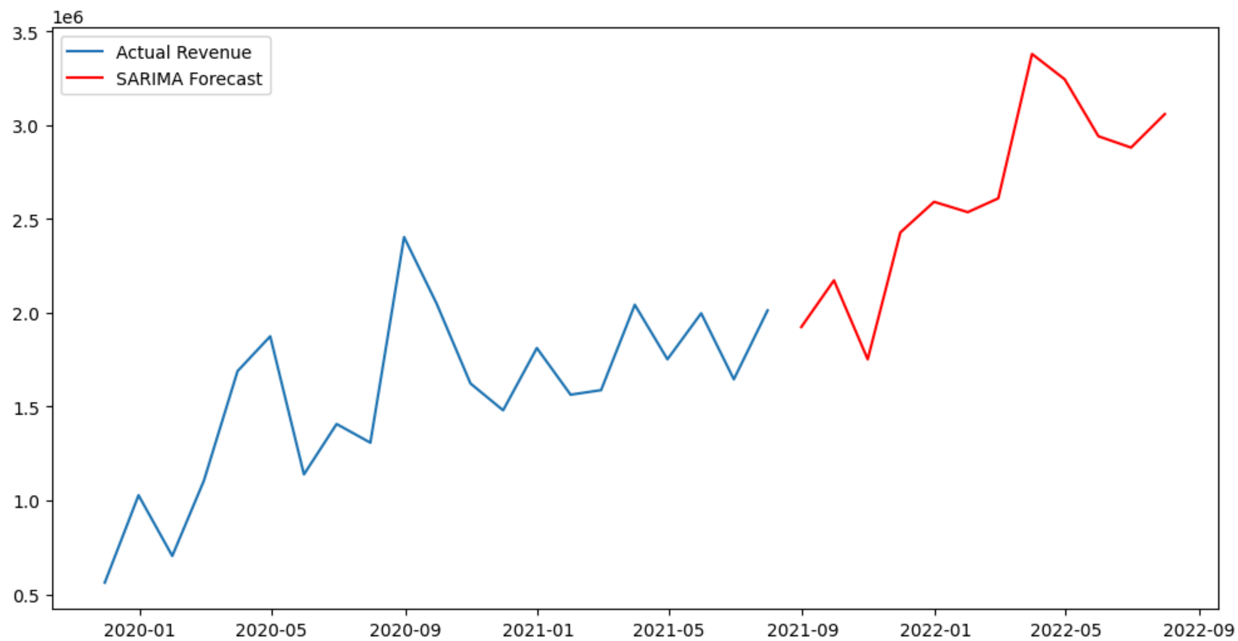
For our baseline model, we chose an AutoRegressive Integrated Moving Average (ARIMA) model with parameters (0, 2, 1). Here, '0' indicates no autoregressive terms—meaning the model does not rely on previous time points. '2' signifies that the data has been differenced twice to make it stationary, addressing trends but not seasonality. '1' denotes a single moving average term, which helps smooth out past noise in the data series. While effective for non-seasonal trends, this model lacked components to handle seasonal variation, which is crucial for accurate forecasting in contexts with seasonal sales cycles.



Parallelly, in our quest to predict claim outcomes, we adopted a Random Forest classifier as our base model for claim predictions.The choice of a Random Forest classifier was due to its robustness in handling binary classifications, ideal for predicting claim outcomes (accepted or rejected). Initially, basic features like CLAIM_TOTAL and Claim_Accepted were included. The model's configuration with 100 trees was intended to balance complexity and performance, achieving an initial accuracy of 0.81 and an F1 score of 0.88, suggesting effective classification at this stage.

3.2 Iterative Model Improvements

Recognizing the importance of seasonality in healthcare data, we refined our forecasting approach with the Seasonal AutoRegressive Integrated Moving Average (SARIMA) model. This advanced model, incorporating both non-seasonal (1, 1, 1) and seasonal (1, 1, 1, 12) components, significantly outperformed the ARIMA model by effectively capturing the seasonal ebbs and flows, thus providing more actionable insights for strategic planning.



In response to the initial success of the Random Forest model in claim prediction, we iteratively enhanced its feature set by incorporating additional variables like DateBilled and AdjudicationDate, which led to remarkable improvements in accuracy (0.96) and F1 score (0.97). Encouraged by these results, we further experimented with a Gradient Boosting Machine (GBM) to refine our predictions. The GBM, while achieving commendable accuracy (0.94) and F1 scores (0.96), ultimately fell short of the enhanced Random Forest model in subsequent robustness tests.

Chapter 4: Results

We first examine the results of the models developed for predicting claim denials and acceptances. Through rigorous testing and analysis, we have evaluated the performance of both the Gradient Boosting Machine (GBM) and Random Forest models. These analyses will delineate how each model performed under various metrics and testing conditions, leading to an informed decision on which model best meets our needs for operational deployment.

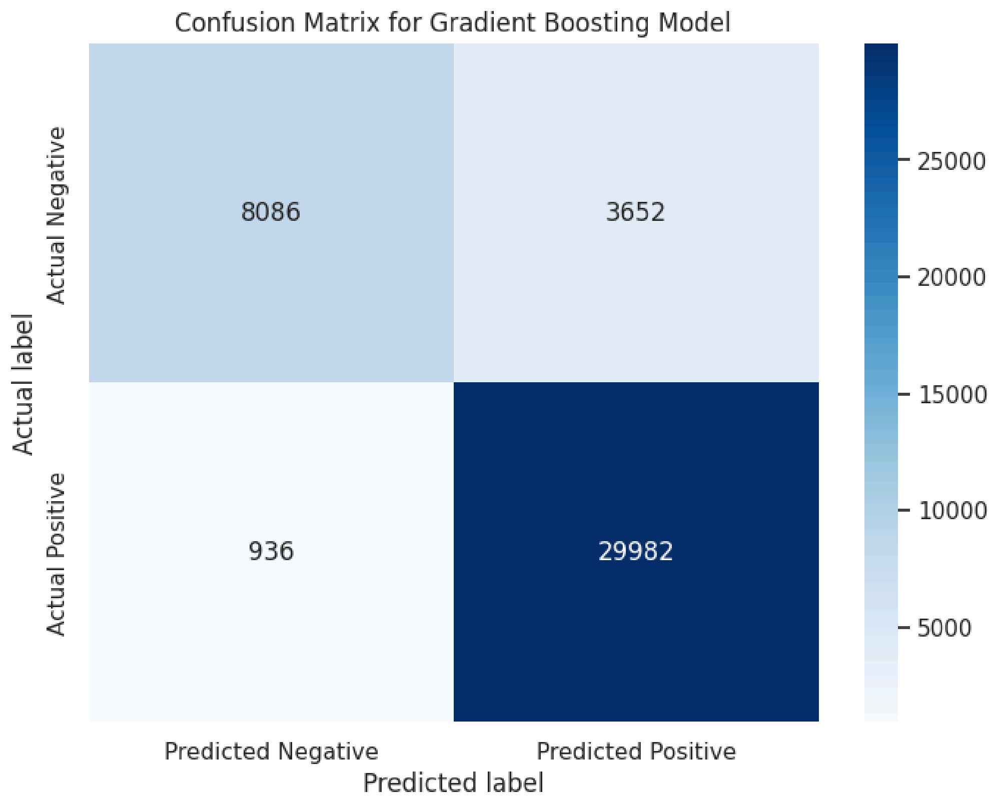
4.1 Random Forest & Gradient Boosting model results

4.1.1 Cross-Validation Analysis

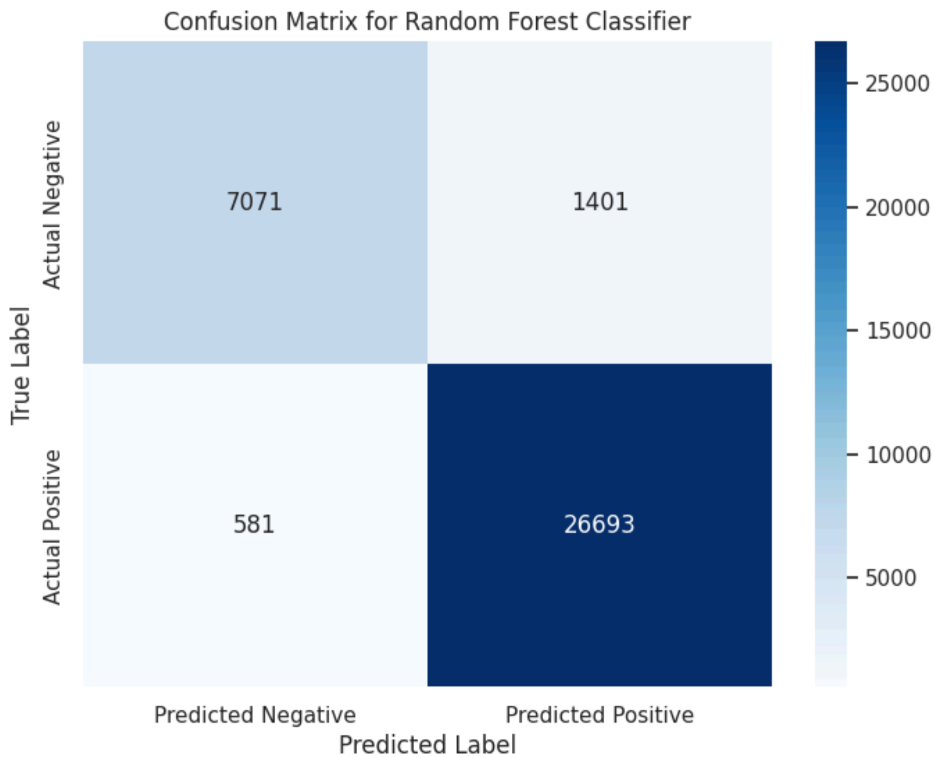
Cross-validation was employed as a primary method to assess the generalization capabilities of our models. The Random Forest model recorded a mean accuracy of 0.81, while the GBM model exhibited a slightly higher mean accuracy of 0.89. Although the GBM showed potential in handling diverse data distributions effectively, the comprehensive metrics evaluated beyond mere cross-validation scores highlighted crucial differences in performance.

4.1.2 Confusion Matrix Analysis

Through confusion matrix analysis, we gained deeper insights into each model's performance:



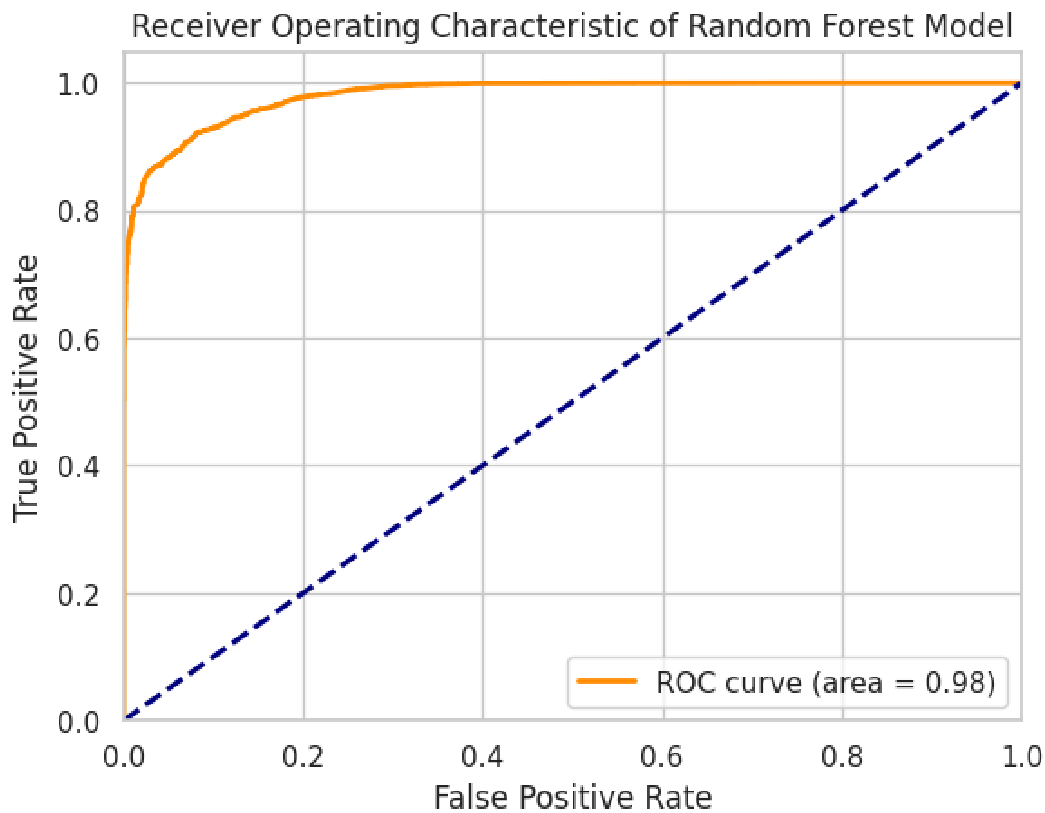
The GBM model demonstrated robust accuracy at 0.9195 and an impressive precision of 0.9860. However, when considering the recall, F1 score, and specificity, it became evident that while the model is highly precise, it does not altogether balance well across all metrics.



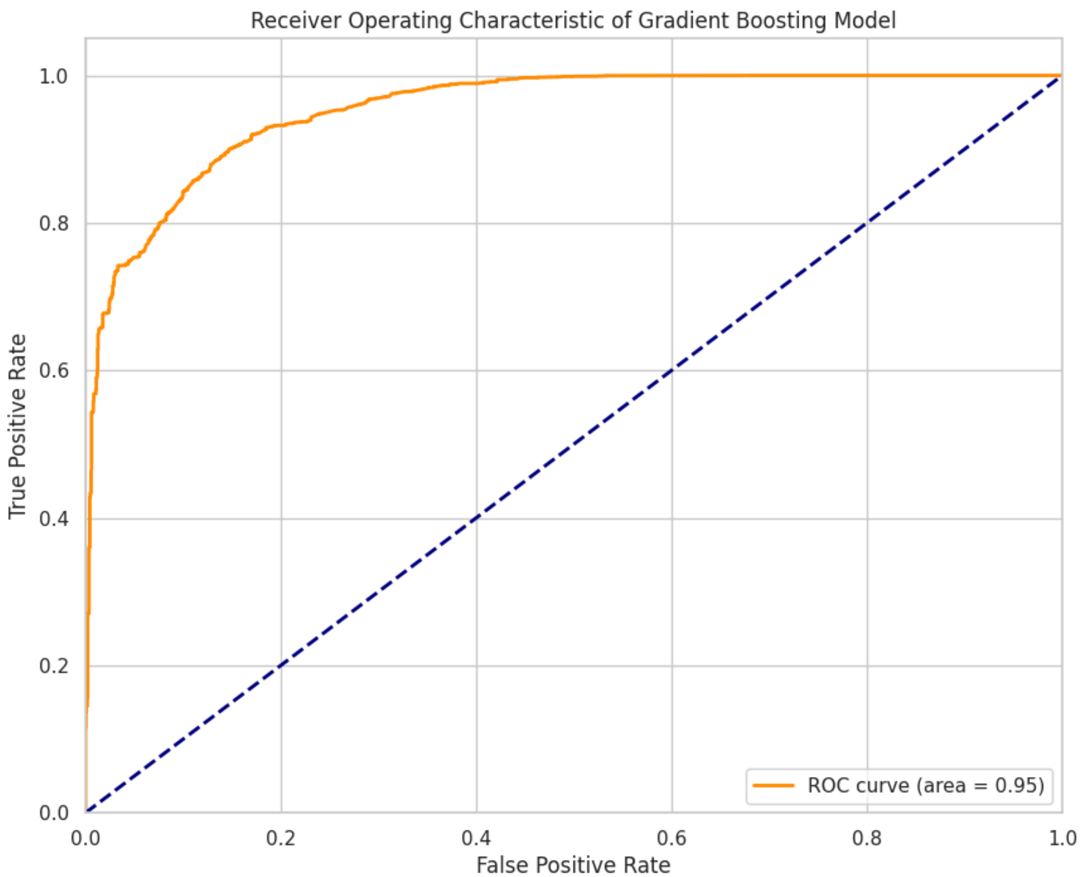
Conversely, the Random Forest model showcased an overall superior balance with an accuracy of 0.9446, an exceptional recall of 0.9787, and an F1 score of 0.9642, indicating its strength in correctly identifying true claims. Despite a slightly higher false positive rate of 0.1654, its performance in critical areas was unmatched.

4.1.3 ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve for the Random Forest model provided a visual and quantitative measure of its ability to discriminate between accepted and denied claims across all decision thresholds. The area under the ROC curve (AUC) is a critical metric representing the model's capability to avoid false classifications. In our analysis, the Random Forest model achieved an AUC of 0.98, indicating a near-perfect classification ability. This high AUC score reaffirms the model's superior accuracy and reliability, showcasing its robustness in maintaining high performance across various decision thresholds.



Similarly, the ROC curve analysis for the Gradient Boosting Machine (GBM) model demonstrated its discriminative capabilities. The GBM model scored an AUC of 0.95, which, while slightly lower than that of the Random Forest, still indicates a strong ability to correctly classify claims. This ROC curve analysis highlights the GBM model's solid performance but also underscores the slightly reduced reliability in avoiding false classifications compared to the Random Forest model.



The detailed comparison and evaluation underline the Random Forest model's superior performance, particularly in its ability to maintain high accuracy and excellent recall, crucial for minimizing the risk of failing to identify valid claims. The Random Forest model's robustness across different decision thresholds, evidenced by its high AUC of 0.98 in ROC curve analysis, confirms its efficacy and reliability. Considering all these factors, the Random Forest model was selected as the better overall option for deployment. It aligns well with our operational goals to ensure no legitimate claim is mistakenly denied, a critical factor in claim management.

4.2 ARIMA & SARIMA Results

In our comparative analysis of the ARIMA and SARIMA models, we employed the Akaike Information Criterion (AIC) and the Dickey-Fuller test to evaluate model performance and the stationarity of residuals. The AIC, which helps in comparing models by balancing goodness of fit and model complexity, revealed that the SARIMA model had a lower AIC value than the ARIMA model, indicating a superior fit to the sales data. This suggests that the SARIMA model is more effective in capturing the seasonal fluctuations inherent in our dataset, a critical aspect for accurate forecasting.

Further validating the model's efficacy, the residuals of the SARIMA model were tested for stationarity using the Dickey-Fuller test. The test results confirmed that the SARIMA model's residuals are stationary, indicating that it adequately accounts for both trends and seasonality, unlike the ARIMA model whose residuals suggested some remaining non-stationarity. This difference underscores the SARIMA model’s capability to provide reliable forecasts that are crucial for strategic planning.

Our efforts to conduct additional tests to further assess and compare the models encountered some challenges, resulting in unsuccessful outcomes. These difficulties highlight the complexities often faced in predictive modeling and underscore the importance of relying on robust, well-established tests like the AIC and Dickey-Fuller test. Such tests are essential for ensuring the models' effectiveness in real-world applications and for making informed decisions about model selection.

Based on the results from the AIC and Dickey-Fuller tests, the SARIMA model emerged as the more appropriate choice for our sales forecasting needs. Its ability to effectively handle both non-seasonal and seasonal variations ensures that it can provide accurate and actionable forecasts, making it an invaluable tool for our business operations. The insights from this rigorous evaluation not only confirm the SARIMA model's suitability for deployment but also highlight the necessity of comprehensive model testing to validate predictive accuracy and reliability.

Chapter 5: Evaluation

Our project was initiated with the dual objectives of enhancing financial management within healthcare billing and optimizing claims handling efficiency. These goals were addressed by developing predictive models for accurate revenue forecasting and effective claim processing. The SARIMA model was developed from an ARIMA baseline to capture seasonal fluctuations in sales data, aligning with our data mining objective to improve sales forecast accuracy. Concurrently, we enhanced a basic Random Forest classifier by incorporating additional features like DateBilled and AdjudicationDate, significantly improving its predictive accuracy for claim outcomes.

5.1 Model Selection and Refinement

Initial model explorations included considerations of various predictive techniques. However, after reviewing past experiences and consulting with our project supervisor, we decided against using the Prophet model due to its previous underperformance in similar applications. An attempt to deploy an LSTM model was also unsuccessful due to technical challenges, reaffirming our decision to focus on refining the SARIMA and Random Forest models. For the Random Forest model, extensive parameter tuning determined that 100 trees optimized performance across multiple metrics, including accuracy and recall.

5.2 Model Testing and Validation

The SARIMA model's reliability was confirmed by the Dickey-Fuller test, which verified the stationarity of the time series data—critical for accurate seasonal forecasting. For the Random Forest model, rigorous validation procedures such as cross-validation and ROC curve analysis were employed. These tests demonstrated the model's robustness and high discriminatory power, with an AUC score reflecting near-perfect classification ability.

5.3 Impact Assessment

The deployment of the SARIMA model has led to more informed decisions regarding inventory management and promotional strategies, significantly enhancing operational efficiency and profitability. The Random Forest model's accuracy in predicting claim outcomes has not only reduced the costs associated with misprocessed claims but has also markedly improved customer satisfaction by ensuring more reliable claim handling.

The systematic evaluation of our predictive models shows that both the SARIMA and Random Forest models have substantially met and exceeded our initial business and data mining objectives. Their implementation has provided strategic benefits in sales forecasting and claims management, underscoring the importance of rigorous model selection, testing, and continuous refinement in predictive analytics.

Chapter 6: Conclusion & Future Work

This project has successfully demonstrated the powerful impact of predictive analytics in healthcare billing and claims management by developing and implementing robust predictive models. The SARIMA and Random Forest models have significantly enhanced decision-making processes, leading to improved operational efficiency and financial management. The SARIMA model adeptly handled seasonal fluctuations in sales data, providing accurate and actionable forecasts, while the Random Forest model reliably predicted claim outcomes, reducing the costs associated with misprocessed claims and enhancing customer satisfaction. Through rigorous evaluation processes, including cross-validation and ROC curve analysis, the models proved their robustness and reliability, exceeding the project's initial objectives and offering valuable insights into the practical applications of advanced analytics in healthcare.

Looking forward, there are several directions for continuing and expanding this work. Integrating additional data sources such as patient demographics and geographical information could refine the models' predictive capabilities further, enabling personalized and location-specific insights. Applying these methodologies to other areas of healthcare, such as resource allocation or patient admission rates forecasting, could broaden the impact of our efforts. Moreover, exploring advanced machine learning techniques like deep learning might uncover deeper patterns in complex healthcare data, potentially leading to more profound insights and innovations.

The development of real-time analytics systems could transform decision-making into a more dynamic process, allowing healthcare providers and insurers to respond swiftly to changing conditions. Continuous research into model optimization and tuning, alongside collaboration with industry experts and stakeholders, would ensure the models' relevance and applicability remain at the forefront of healthcare analytics. Such collaborative efforts would also provide critical feedback necessary for guiding future enhancements.

The achievements of this project highlight the transformative potential of data science in healthcare, showing how targeted analytics can lead to substantial improvements in service efficiency and patient care. As we continue to innovate and build on this foundation, the prospects for further enhancing healthcare services are both promising and exciting. By pushing the boundaries of what is possible with data analytics, we can continue to contribute to more effective healthcare operations and better patient outcomes.

References

1. N. Beheshti, "Guide to Confusion Matrices & Classification Performance Metrics," Towards Data Science, [Online]. Available: https://towardsdatascience.com/guide-to-confusion-matrices-classification-performance-metrics-a0ebfc08408e. [Accessed: Jan. 15, 2024].
2. N. Beheshti, "Random Forest Regression," Towards Data Science, [Online]. Available: https://towardsdatascience.com/random-forest-regression-5f605132d19d. [Accessed: Feb. 28, 2024].
3. N. Beheshti, "Cross-Validation and Grid Search," Towards Data Science, [Online]. Available: https://towardsdatascience.com/cross-validation-and-grid-search-efa64b127c1b. [Accessed: March 3, 2024].
4. A. Mesri, "Data Science Project: Sales Forecasting with ARIMA," Medium, [Online]. Available: https://medium.com/@alpmesri/data-science-project-sales-forecasting-with-arima-e377af0ffdaa. [Accessed: Jan. 10, 2024].
5. D. S. S. Sandha, "Tuning ARIMA for Forecasting: An Easy Approach in Python," Medium, [Online]. Available: https://medium.com/@sandha.iitr/tuning-arima-for-forecasting-an-easy-approach-in-python-5f40d55184c4. [Accessed: Feb. 20, 2024].
6. W. Koehrsen, "Hyperparameter Tuning the Random Forest in Python Using Scikit-Learn," Towards Data Science, [Online]. Available:

https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74. [Accessed: March 25, 2024].

7. "What is a CPT® code?" American Medical Association, [Online]. Available: https://www.ama-assn.org/practice-management/cpt/cpt-overview-and-code-approval. [Accessed: April 9, 2024].
8. "List of Codes Effective January 1, 2023," Centers for Medicare & Medicaid Services, [Online]. Available: https://www.cms.gov/license/ama?file=/files/zip/list-codes-effective-january-1-2023-published-december-1-2022.zip. [Accessed: Jan. 20, 2024].
9. "Receiver Operating Characteristic (ROC) with cross-validation," scikit-learn, [Online]. Available: https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html. [Accessed: Feb. 25, 2024].
10. "sklearn.ensemble.RandomForestClassifier," scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. [Accessed: March 10, 2024].

Acknowledgements

We extend our deepest gratitude to our project supervisor, Dr. Alex, for his invaluable guidance and expertise throughout the duration of this project. His support was crucial in helping us navigate the complexities of our research and in shaping the direction of our work.

We are also immensely thankful to our course professor, Muazzam Siddiqui, whose teachings and insights have profoundly influenced our approach to this research.

Professor Siddiqui's dedication to academic excellence and his commitment to nurturing our potential have been instrumental in our success.

Additionally, we would like to express our appreciation to our teaching assistant, Kanin Bender, for his assistance and encouragement. Kanin's readiness to help and his attention to detail have greatly enhanced our learning experience and contributed significantly to the completion of this project.