Project Report: Predictive Modeling for Healthcare Facility Analysis

Anay Halwasiya, Ashwin Kumar, and Jatin Jain

Introduction

The healthcare industry is a vital sector where even minor issues—such as lapses in infrastructure, quality of care, or resource allocation—can have significant consequences for communities and populations. Regular audits are essential for maintaining compliance, upholding quality standards, and identifying gaps in operations. However, audit flags often highlight deeper issues like compliance violations, financial instability, or operational inefficiencies, which can harm a hospital's reputation, erode patient trust, and result in costly penalties or corrective actions.

This project addresses these challenges by developing a machine learning-based predictive model to analyze healthcare facility performance metrics and identify audit risks. The model provides hospitals with actionable insights to predict the likelihood of audits and highlight key risk factors. With this information, hospital administrators can enhance compliance, optimize resource allocation, and implement targeted operational improvements to minimize audit risks and improve overall efficiency.

Given the strict regulatory oversight and resource constraints faced by healthcare systems, addressing audit risks is critical to ensuring seamless operations and maintaining public trust. This project empowers hospitals to proactively identify vulnerabilities, mitigate disruptions, and enhance decision-making through data-driven insights. By fostering compliance and operational efficiency, this work contributes to building more resilient and effective healthcare systems.

Methods

Algorithms Used

To achieve robust predictions and actionable insights, we implemented a combination of Logistic Regression, Random Forest, and XGBoost algorithms with varying regularization methods. They were subsequently integrated into an ensemble model using a Voting Classifier. Logistic Regression was selected as the baseline due to its simplicity and interpretability, offering a clear understanding of the relationships between features and the target variable. Random Forest, with regularization and its ensemble of decision trees, was selected for its ability to handle non-linear relationships and high-dimensional data. XGBoost, a gradient-boosting algorithm, was chosen for its high

predictive power and advanced regularization techniques, which enhance model generalization and robustness. The Voting Classifier aggregated predictions from these models, leveraging their strengths to improve accuracy and reliability. A soft voting approach was adopted to combine the probabilistic outputs of each model, further enhancing the ensemble's performance.

Preprocessing Steps

Data preprocessing was a critical component of this project, ensuring that the dataset was clean, consistent, and ready for analysis. Missing data in key features such as bed availability (BED_AVL) and licensure (BED_LIC) were imputed with default values to maintain dataset integrity. Feature engineering played a significant role in capturing relevant patterns; for instance, a new variable representing bed utilization (Bed_Utilization) was derived to quantify resource efficiency. Categorical variables, such as facility types (TYPE_CARE), were consolidated to reduce redundancy, while geographic and temporal features were transformed into formats suitable for analysis. Continuous features were normalized to ensure uniform contributions to the models, and stratified sampling was employed during the train-test split to preserve class distributions in the target variable (AUDIT_IND).

Code Overview

The implementation of this project was structured into modular code for preprocessing, modeling, and evaluation. Key libraries used include pandas and numpy for data manipulation, scikit-learn for model development, and XGBoost for gradient boosting. The Voting Classifier was implemented using scikit-learn's ensemble module. A dedicated script handled all preprocessing steps, ensuring data consistency and compatibility with machine learning models. Individual models were trained using their respective libraries. The Voting Classifier integrated these models, leveraging their strengths to make robust predictions. Regularization was used in the models to reduce overfitting and make the models generalize better. Performance metrics, including log loss and ROC-AUC, were calculated to assess model effectiveness. Results were visualized using matplotlib and seaborn.

Dataset

The dataset selected, California Annual Hospital Financial Data from Data.gov (https://catalog.data.gov/dataset/hospital-annual-financial-data-selected-data-pivot-tables-92074), contained healthcare facility information, including identifiers (FAC_NO, FAC_NAME), geographic details (COUNTY, CITY), care types (TYPE_CARE), and metrics such as bed availability (BED_AVL) and licensure (BED_LIC). The data required

significant cleaning and preprocessing to address inconsistencies in categorical variables and missing values.

Evaluation and Analysis

Study Design

The evaluation of our models focused on assessing their predictive performance and interpretability using a variety of quantitative metrics and visualizations. The dataset was divided into training and testing subsets, ensuring the models were evaluated on unseen data to measure their generalizability. A stratified train-test split was employed to maintain the distribution of the target variable across both subsets.

Evaluation Metrics

To evaluate the model's performance comprehensively, the following metrics were utilized:

- **ROC-AUC**: Assesses the model's ability to distinguish between classes, with a higher score indicating better discrimination.
- **Log Loss**: Evaluates the quality of probabilistic predictions, penalizing overconfident and incorrect predictions to ensure robust model reliability.

Results

1. **Baseline Model (Logistic Regression)**: Achieved moderate performance based on ROC-AUC (0.668) and Log Loss(0.626). Regularization was applied to prevent overfitting and improve model generalization. However, the model's inability to capture non-linear relationships limited its predictive capability.

2. **Random Forest**: Demonstrated improved performance over the baseline model, achieving better ROC-AUC (0.802) and Log Loss (0.505) due to its ability to model complex feature interactions. Regularization techniques, such as limiting tree depth and tuning minimum samples per leaf, further enhanced its performance and prevented overfitting.

3. **XGBoost**: Outperformed the Random Forest, achieving the highest ROC-AUC (0.84) and lowest Log Loss (0.477) among individual models. Regularization techniques, including L1 and L2 penalties, were employed to fine-tune the model

and optimize predictive accuracy, further improving its ability to generalize.

4. **Voting Classifier**: Combined the strengths of individual models and achieved the best overall performance. Regularization was incorporated into the underlying models, which ensured balanced improvements in ROC-AUC (0.871) and Log Loss (0.483). The ensemble model effectively leveraged diverse algorithms to enhance predictive stability and accuracy.

Key Insights

- Most of the facilities had an audit probability between 0.1 and 0.4, with the observed mode at around 0.2 i.e. most facilities had an approximately 20% chance of an audit
- Porterville State Hospital had the highest probability (around 99%) of being audited followed by Kaiser Foundation Hospital - Santa Rosa (around 85%) which shows a large gap between the first and second facilities in the Top 10 facilities.
- Stanford University Hospital had the lowest audit probability of around 4%.
- For the 10 hospitals with the highest audit probability, Managed Care Contract Utilization and Licensed Beds and Utilization by Type of Care had the highest probability values among the areas.
- For the 10 hospitals with the lowest audit probability, the audit probabilities varied between features. Still, Capitation Premium Revenue had the highest probability values showing a contrast with the top 10 hospitals in features that are likely to lead to an audit.

Related Work

Hospital capacity strain has long been a concern, with major implications for patient care and operational efficiency. One study reviews strategies hospitals use to manage high-bed occupancy, emphasizing the role of predictive models and resource optimization, aligning with our project's objectives. [3] Another study links high occupancy to negative patient outcomes, highlighting the need for proactive management tools. [4] While direct studies on audit prediction are limited, research in compliance risk modeling and healthcare resource optimization offers relevant methods. These studies demonstrate the value of machine learning in identifying high-risk entities, similar to our approach for predicting audit likelihood. Additionally, research on predictive analytics in healthcare emphasizes the need for interpretable models, which aligns with our focus on providing actionable insights. Our project addresses the gap between audit prediction and operational efficiency, contributing to data-driven decision-making in healthcare.

Discussion and Conclusion

Interpretation of Results

The predictive models developed in this project effectively predicted the likelihood of a healthcare facility being audited and identified key risk areas. Notably, the ensemble approach using the Voting Classifier showed significant improvements over the baseline Logistic Regression model, as evidenced by higher AUC and lower log loss scores (Appendix Fig. 1). This underscores the effectiveness of ensemble methods in handling complex healthcare data with multiple variables and relationships.

We also found that most facilities had around a 20% audit probability, with 8 out of the 10 facilities with the highest audit probabilities located in or around the San Francisco Bay Area, where Kaiser has a strong presence. The Top 10 facilities, compared to the Bottom 10 (Appendix Figs. 5 and 6), revealed distinct patterns in audit risk, with Managed Care being the dominant risk factor for the Top 10, while the Bottom 10 had varied risks across areas like Capitation Premium Revenue, Patient Revenue Information, and Direct Expenses By Cost Center.

Ultimately, the model's ability to predict audit likelihood enables facilities to take proactive measures. Insights derived from the model can inform decisions on staffing, financial practices, and operational adjustments, helping hospitals mitigate risks and avoid reputational damage and costly penalties for compliance violations.

Key Learnings

- **Overfitting and Model Generalization**: A significant challenge observed during the project was overfitting, as indicated by the noticeable gap between training and testing scores across all models for both metrics. While the models excelled in training data, their performance on unseen data was less robust. To counter this issue, we applied techniques such as regularization and hyperparameter tuning, which helped to control model complexity and improve generalization. These strategies underscore the importance of balancing model flexibility and predictive power, particularly in complex healthcare datasets.

- **Effectiveness of Ensemble Methods**: Ensemble methods, especially the Voting Classifier, demonstrated their ability to enhance predictive performance by effectively balancing the trade-offs between bias and variance. By combining the strengths of multiple algorithms, these methods captured diverse patterns in the data, resulting in more accurate and reliable predictions. This proves especially

valuable in healthcare settings, where precision and stability in predictions are critical for decision-making and resource allocation.

Future Work

While the current model provides valuable insights, several areas could be explored to enhance its effectiveness and applicability further. A limitation of the current model is that the dataset used is specific to California, which may limit its ability to generalize to other regions. Expanding the dataset to include data from hospitals across the United States, or even internationally, could enhance the model's robustness and generalizability. Incorporating a more diverse set of healthcare facilities, with varying regulatory environments, operational structures, and patient populations, would help capture a broader range of factors that contribute to audit risks. This expansion would likely improve the model's predictive accuracy and its applicability to hospitals outside California. Another area for future work involves integrating the model with real-time data streams. By continuously updating the model with fresh operational data, hospitals could monitor audit risks in real time, allowing for dynamic adjustments as new risks emerge. This would provide hospital administrators with up-to-the-minute insights, enabling quicker and more proactive responses to potential issues, rather than relying on periodic audits.

Conclusion

In conclusion, this project effectively demonstrates the potential of machine learning to mitigate audit risks in hospitals by predicting audit probabilities and pinpointing areas for improvement. Using a dataset from California, we explored the capabilities of XGBoost, Random Forests, and Ensemble Methods to capture the data's complexities, leveraging feature engineering and model selection to predict future outcomes of audits. Despite challenges such as overfitting and data dimensionality, we addressed these issues with regularization techniques and dimensionality reduction.

We were also mindful of the interpretability of our models, selecting metrics that strike a balance between accessibility and rigor for practical application. Moving forward, expanding the dataset, integrating advanced modeling techniques, and incorporating real-time data present exciting opportunities for future research in this domain.

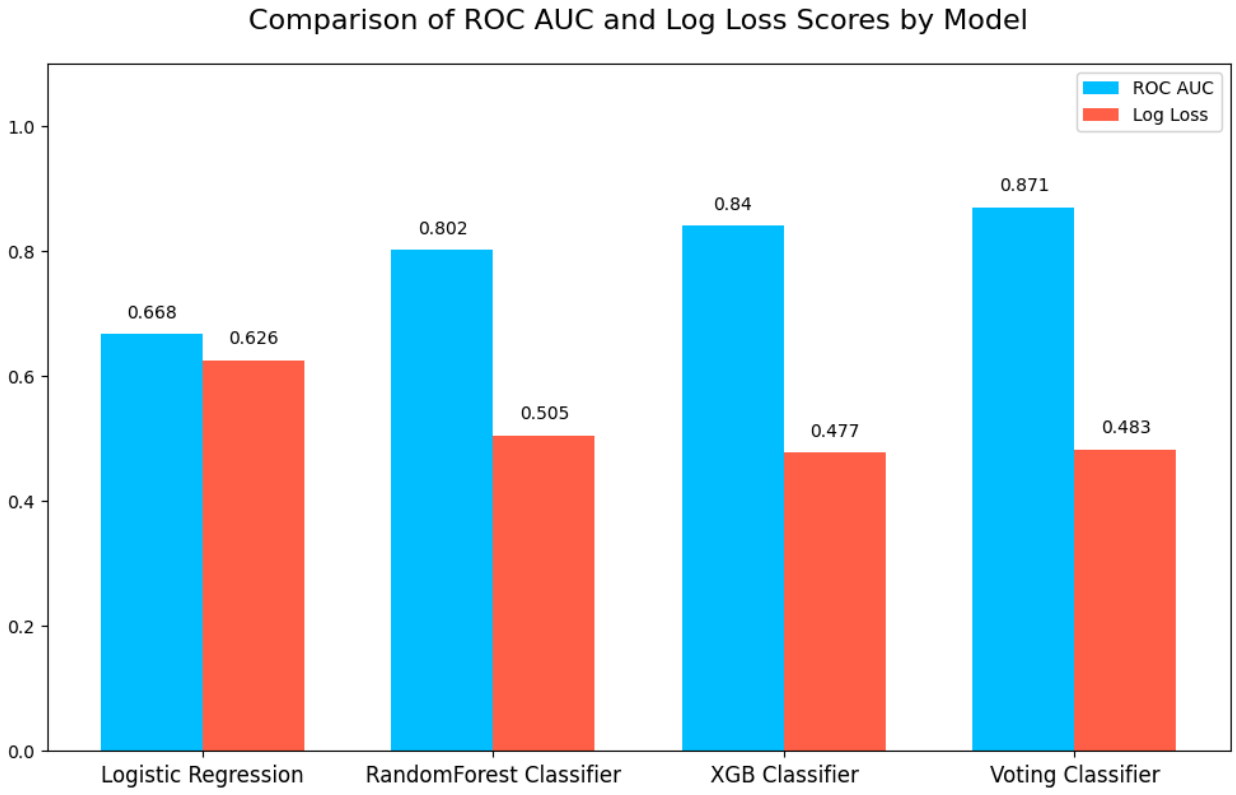
Ultimately, this project contributes to the development of a data-driven tool that can enhance decision-making, optimize resource allocation, and improve compliance within hospitals, fostering more efficient and resilient healthcare systems.

References

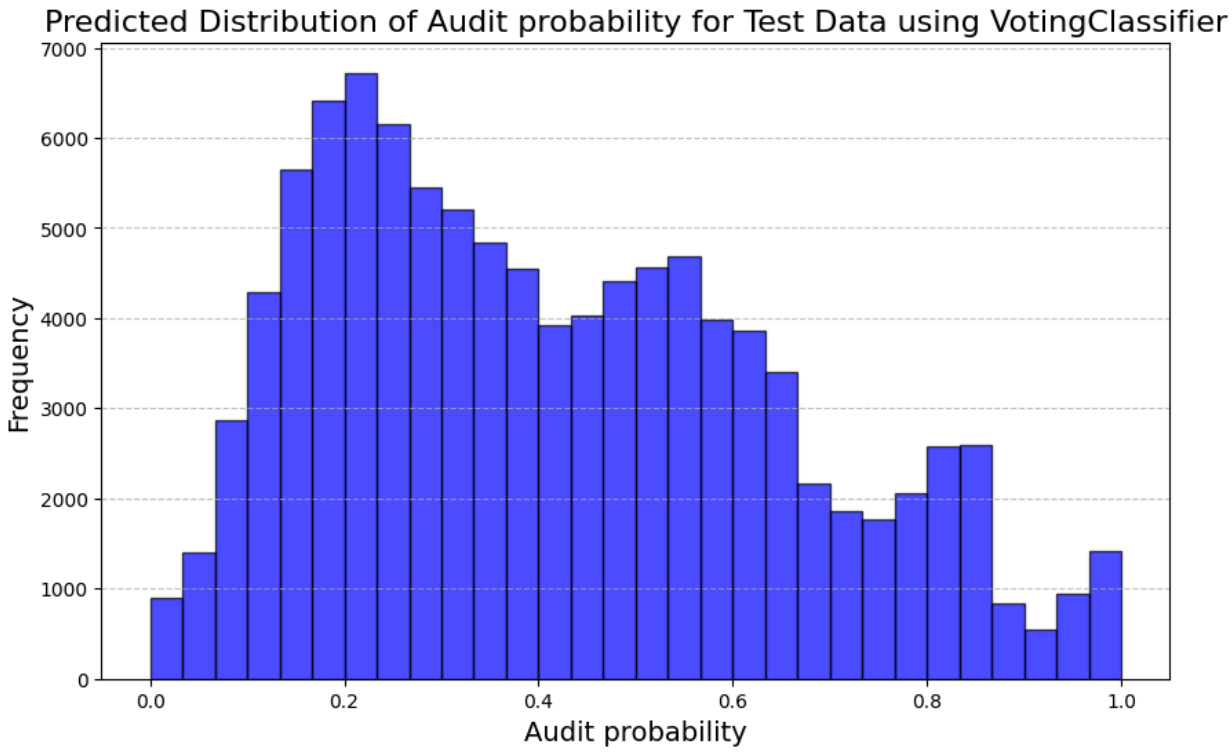
1. Scikit-learn documentation: https://scikit-learn.org
2. XGBoost documentation: https://xgboost.readthedocs.io
3. Arogyaswamy S, Vukovic N, Keniston A, Apgar S, Bowden K, Kantor MA, Diaz M, McBeth L, Burden M. The Impact of Hospital Capacity Strain: a Qualitative Analysis of Experience and Solutions at 13 Academic Medical Centers. J Gen Intern Med. 2022 May;37(6):1463-1474. doi: 10.1007/s11606-021-07106-8. Epub 2021 Dec 13. PMID: 34902096; PMCID: PMC8667526
4. Bosque-Mercader, L., Siciliani, L. The association between bed occupancy rates and hospital quality in the English National Health Service. *Eur J Health Econ* 24, 209–236 (2023). https://doi.org/10.1007/s10198-022-01464-8
5. ChatGPT was used for general purposes in the making of the project report

Appendix

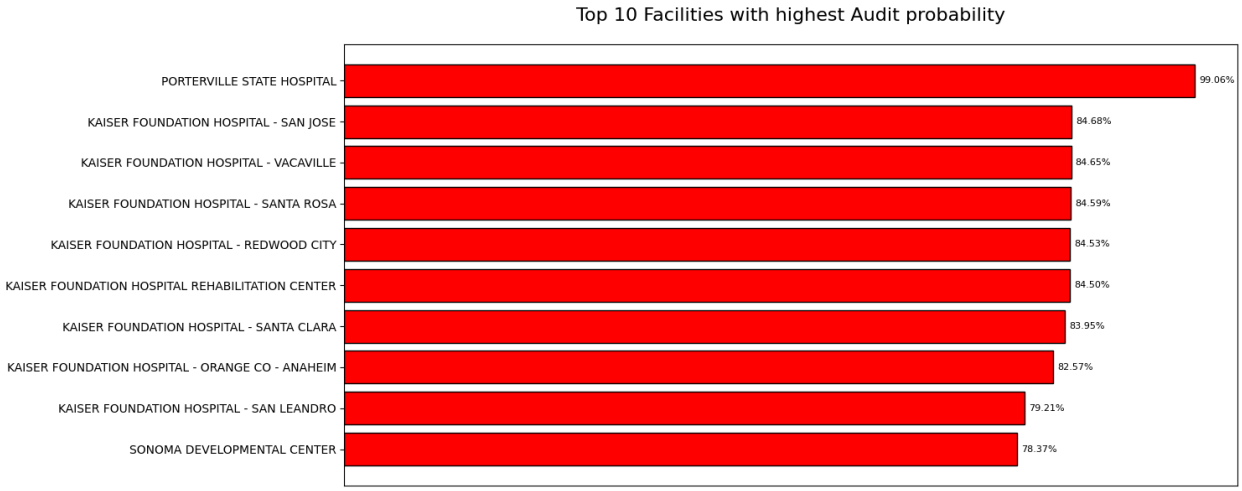
1. Final Comparison of Metrics for All Models



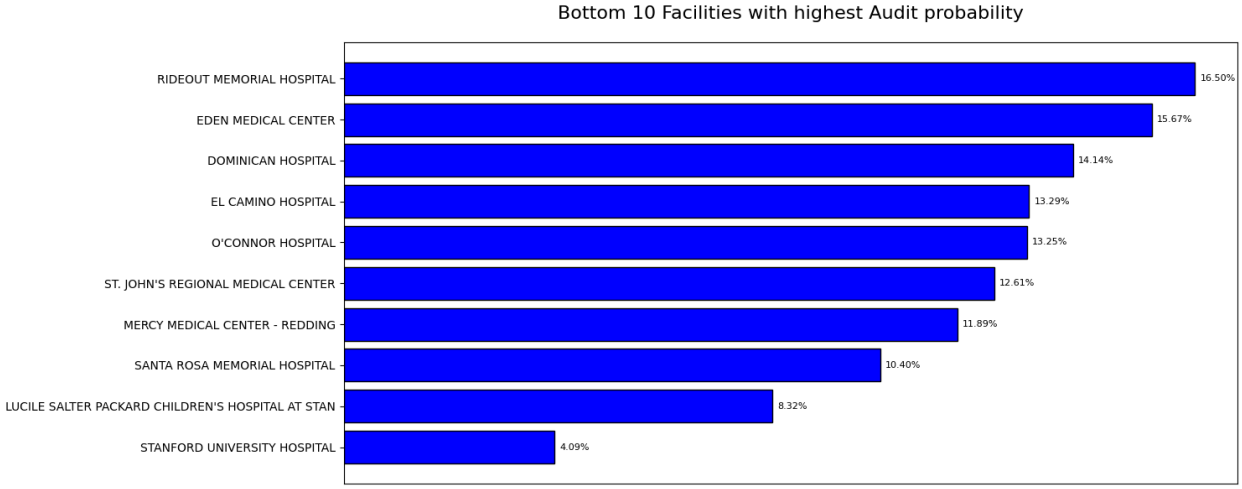
2. Figure showing the Distribution of Audit Probabilities



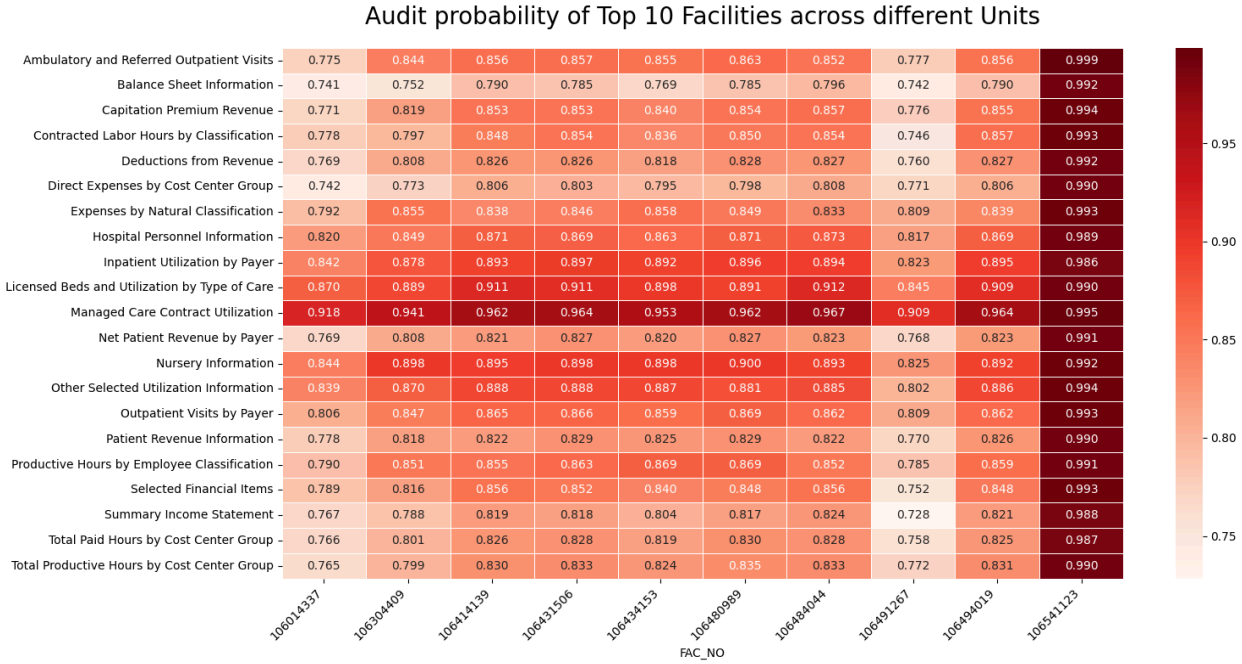
3. Figure showing the Top 10 Facilities i.e. facilities with the highest probability of audits



4. Figure showing the Bottom 10 facilities i.e. facilities with the lowest probability of audits



5. Heatmap showing the probabilities of audits in different units of the facility for Top 10 Facilities



6. Heatmap showing the probabilities of audits in different units of the facility for Bottom 10 Facilities