Banking Customer Churn Prediction and Analysis

INTRODUCTION

The objective of this project is to develop predictive models for customer churn in the banking industry. Maintaining a loyal customer base is a crucial aspect of business success, as the cost of acquiring new customers are typically higher than retaining existing ones. By identifying the factors that contribute to customer churn, banks can take proactive measures to retain customers and improve overall customer satisfaction. This project employs binary classification techniques, including logistic regression, random forest and XGBoost, to predict customer churn and determine the most influential predictors. The model's performance is evaluated using standard metrics such as accuracy, precision and recall, and the model with the lowest total error will be selected as the optimal model.

EXPLORATORY DATA ANALYSIS

The dataset comprises of 10,000 rows and 14 columns, with a churn rate of 20%. There were no missing values in the dataset. Our target variable was "exited", and we dropped three unnecessary features, namely row number, customer id and surname. The remaining 10 columns were credit score, geography, gender, age, tenure, number of products, whether the customer had a credit card or not, whether the customer was an active member or not, and estimated salary. Among these attributes, 2 were categorical variables and the rest were numerical.

To better understand the dataset, we performed exploratory data analysis and visualizations. Our analysis revealed that customer who did not have a credit card were more likely to churn compared to those who had a credit card. Furthermore, we discovered that customers from Germany were more likely to churn compared to customers from France and Spain.

FEATURE ENGINEERING

Featuring engineering is a crucial component in the machine learning pipeline, as it can significantly impact the performance of the model. The process involves transforming data into forms that better relate to the underlying target to be learned. Effective feature engineering techniques can enhance the value of existing data and improve the performance of machine learning models.

To optimize our dataset, we applied several feature engineering techniques, including categorical encoding, Principal Component Analysis(PCA), new feature creation, and feature splitting and scaling. We converted the categorical variables of geography and gender into numerical variables. For geography, we utilized one-hot encoding as there was no sequence or order. Meanwhile, we applied label encoding to transform the gender attribute into a numerical variable.

We also attempted PCA to reduce the number of features, but found that removing redundant information did not improve model performance. We identified outliers in the age feature, particularly for the Exit 0 group, and explored the creation of a new feature that indicated whether a data point was an outlier or not with respect to age. Despite initial belief that these outliers could be predictive of customer churn, validation recall didn't improve, and we abandoned these two

methods.

After experimenting with PCA and new feature creation, we moved forward with the dataset as is. We split the dataset into a training and test set with a 20% of test size. The training set was further divided into training and validation sets with a split size of 20%. We prepared the data for modeling by performing data normalization and standardization. Normalization was applied to bring numeric values to a common scale without distorting differences in the range of values or losing information. Standardization was performed to transform the values to have a mean of 0 and a standard deviation of 1.

METHODS

Customer churn prediction was performed using three different models: logistic regression, random forest, and XGBoost. Each model was evaluated with three variations to identify the optimal model configuration.

Logistic regression is a popular model for binary classification that estimates the probability of an outcome being one class versus the other. The complexity of the model depends on the number of features used and the size of the dataset. We evaluated three variations of the logistic regression models: the base model, the model with an L1 penalty, and the model with an L2 penalty. The tuned model with an L1 penalty outperformed the other two models in all evaluation metrics such as accuracy, recall, precision and f1-score for both the training and validation sets. Recall is a metric to evaluate how well the model was able to identify the customer who actually churned. Achieving high recall is desirable because it means that the bank can take appropriate actions to retain customers who are likely to churn, ultimately reducing customer attrition rates and improving overall customer satisfaction. The validation recall for the tuned model with L1 penalty was 58%, higher than the recall in both the based model and the tuned model with L2 penalty, which were only 51%. The L1 regularization performed better than L2 in feature selection.

Random forest is an ensemble learning method that combines multiple decision trees to create a more accurate and robust model. It is a popular choice for handling missing data and outliers. The

complexity of the modeling approach depends on the number of trees in the forest, the depth of each tree, and the number of features used to split each node. Increasing any of these parameters could increase the model's complexity and potentially lead to overfitting. However, random forest has built-in mechanism such as bagging to reduce overfitting and improve model performance.

To optimize our model's performance, we focused on two critical hyperparameters that control the size and depth of the tree ensemble: the number of trees(n_estimator) and maximum depth of the trees (max_depth). We used a grid search to vary these hyperparameters over a range of values to find the best combination that yielded the highest model performance. The base model, which had no hyperparameters tuning, achieved an almost perfect training accuracy of 0.99. We then used grid search to find the best parameters, with n_estimators =80 and max_depth =10. Finally, we used smaller trees with n_estimators= 10 and max_depth =3 to access the effect on model's performance. The model with smaller trees was better at identifying true positives (churned) but struggled to identify all the positive cases. The second variation of random forest outperformed the other two, achieving high recall (86%) in validation sets. This result demonstrates that random forest with tuned hyperparameters can effectively identify customers who are likely to churn.

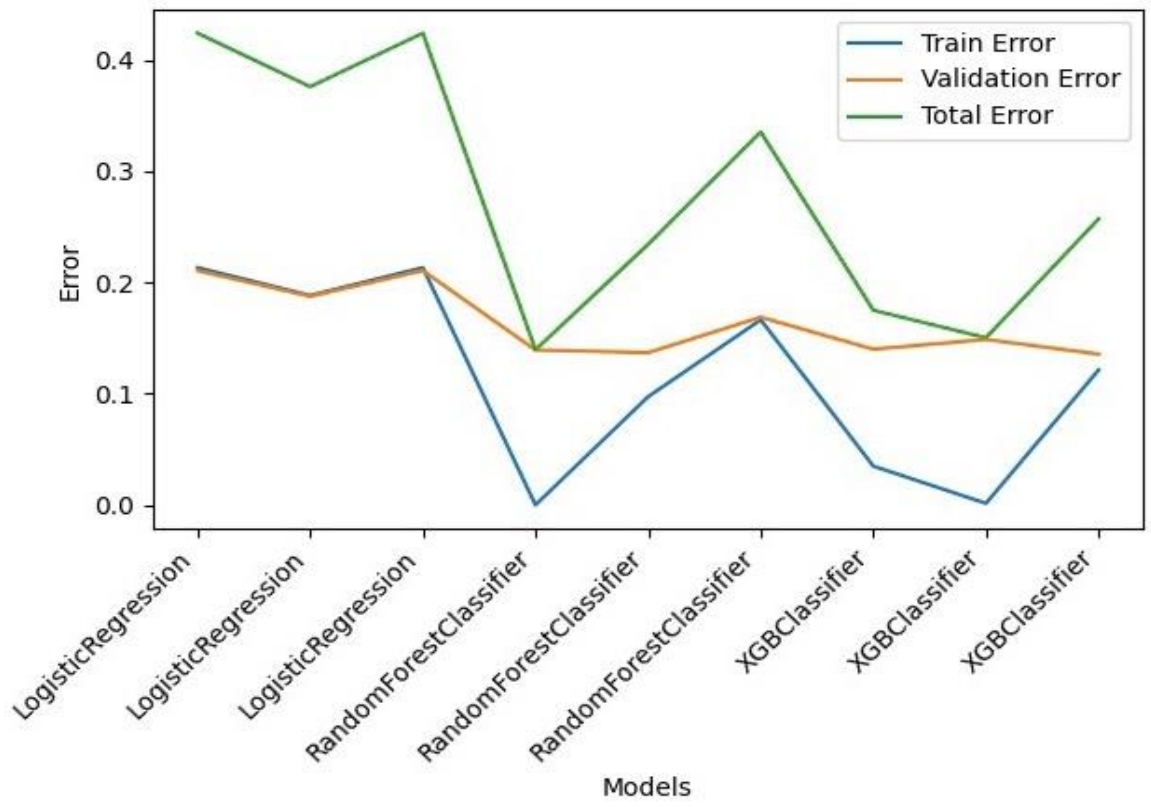
XGBoost is a gradient boosting algorithm that uses an ensemble of decision trees. Unlike random forest, each decision tress is built sequentially to correct the errors of the previous tree. We evaluated three variations of the XGBoost models: XGBoost without hyperparameters tuning, XGBoost with decreased learning rate to 0.1 and max depth to 5, and XGBoost with decreased n_estimators to 140 and max_depths to 3. All three XGBoost models achieved a validation recall of 72%. The third variation achieved the highest accuracy (86%) and precision (82%) on the validation set.

To address the imbalance in the dataset with 80% of data in class 0 (retained) and 20% of data in class 1 (churned), we employed Synthetic Minority Over-sampling Technique(SMOTE) on the third variation of the XGBoost model. SMOTE generates synthetic samples for the minority class by interpolating between the minority samples.

RESULTS

The winning model for predicting customer churn in the banking industry was XGBoost variation 3, which had the lowest validation error among the nine models tested. This model had a learning rate of 0.1, n_estimator of 140, max_depth of 3, min_child_weight of 1, gamma of 0 and subsample of 0.8. To evaluate the model's performance, we analyzed the bias-variance trade-off. Bias measures how far off in general the model's predictions are from the correct value. Variance is how much the predictions for a given point vary between realizations of the model. A model with high bias tends to have training error (underfitting) and a model with high variance tends to have high validation error (overfitting). Therefore, it is crucial to find the optimal balance between bias and variance to avoid overfitting or underfitting the data.

The graph of training, validation and total errors of for the models showed that logistic regression had the highest validation error due to its simplicity, which resulted in high bias and underfitting. The random forest and XGBoost models had lower validation error, indicating better data modeling. However, they also showed some variance among their variations, suggesting overfitting to some extent.



To evaluate the selected model's predictive accuracy, we used metrics of accuracy, precision, recall and F1 score. The model achieved an accuracy of over 86%, while the precision was 75%. However, the recall was low at only 47%. A low recall means that the model was not efficiently identifying customers who churn, leading to potential revenue loss for banks. Given the imbalanced dataset, with class 0 being much more frequent than class 1, we applied SMOTE to balance the dataset, which led to a notable improvement in recall to 67% from the original 47%. However, this came at a cost, as the model's precisions decreased due to an increase in false positives.

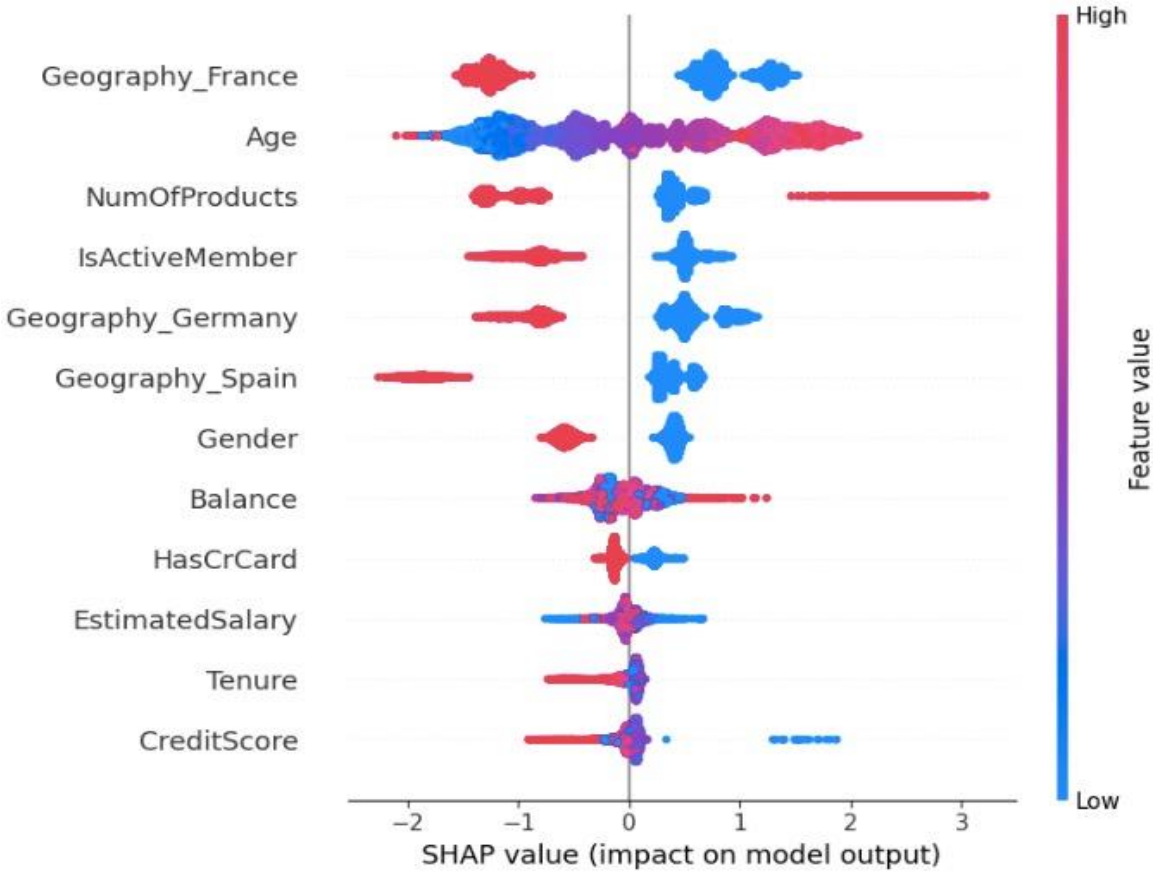
DISCUSSION

1) Feature Importance

We used the SHAP value to evaluate the importance of features in predicting banking customer churn. The SHAP summary plot provides valuable information on the most important features and their range of effects over the dataset. Positive SHAP values are indicative of churning, while

negative SHAP values are indicative of staying at the same bank. As demonstrated by the color bar, higher values are represented in red, while lower values are represented in blue.

Our analysis showed that the top 10 features that impact customer churn are geography France, age, number of products, whether the customer is an active member or not, geography Germany, geography Spain, gender, balance, whether the customer has a credit card or not and estimated salary. This analysis also revealed that a higher value of geography France leads to lower chance of churn, while a higher value of gender(male) and an active membership status leads to a lower chance of churn.



2) Robustness and Bias Analysis

We conducted a robustness and bias analysis to evaluate the impact of removing age and gender variables from the model, given their protected status and potential ethical implications. It is worth noting that inclusion of gender and age in a model do not necessarily imply bias. Our analysis

indicated that age and gender variables are crucial for accurately predicting customer churn. When these variables were removed, the recall in the test set decreased from 67% to 55%, indicating a significant decrease in model performance.

3) Potential Risks

Model drift is a critical concern for any predictive model in production. As our model is continuously receiving new data, it is possible that the new data may have a different probability distribution than the one used during model training. This could lead to a drop in model performance and an increase in false negatives. Additionally, the use of the model may require access to sensitive customer data, raising privacy concerns. To mitigate these risks, we need to monitor the model's performance regularly and establish monitoring thresholds to alert us when the model's behavior deviates significantly from the expected values. Key aspects to monitor include data distribution, feature importance, and model performance, which can be evaluated using performance metrics such as accuracy, precision and recall as well as business metrics such as customer churn rate and uptake in promotions.

By setting appropriate monitoring thresholds, we can detect and respond to any issues promptly, minimizing the risks associated with model drift. For example, if the acceptable range for customers churn rate is between 0-5%, a threshold of 5% could be set as the yellow flag indicating that errors should be tracked closely. A threshold of 10% could be set as the red flag indicating that the model should be pulled out of production.

Once model drift is detected, we can adopt different strategies to handle it. One strategy is to retrain or adapt the model. If the drift is due to changes in data distribution, the model can be retrained or adapted by adjusting model parameters such as training weights to account for the changes in information carried by the data features. Another strategy is to update training data with new data that carries current information about the relationship between the input and output data, and then retrain the model. The frequency of model retraining depends on the rate of data change and the specific requirements of the use case.

FUTURE WORK

While the XGBoost model was effective in predicting customer churn, further improvements can be explored through more extensive tuning of hyperparameters such as the subsample rate and number of tress. This can be achieved by using more exhaustive search techniques for hyperparameters and evaluating performance through cross-validation. Furthermore, deep analysis of potential bias should be conducted, and steps should be taken to address this issue.

Additionally, investigating the reasons behind the most significant features in predicting customer churn, such as geography, age, and number of products, would provide valuable insights into customer behaviors and enable banks to take proactive measure to retain customers.

CONCLUSION

Predicting customer churn is crucial for banks to retain their customers and take proactive measures to prevent them from leaving. In this project, the XGBoost model performed the best, and by utilizing SMOTE to balance the dataset, we were able to improve recall and enhance the reliability of our predictions. Our analysis identified geography, age and the number of products a customer has as the most significant features in predicting customer churn. However, potential bias associated with protected variables such as age and gender should be considered when interpreting the model's predictions. Regular monitoring of the model is necessary to mitigate the risks of model drift. Therefore, maintaining a culture of ongoing monitoring and improvement, and to work collaboratively across teams is imperative to ensure that models remain accurate, effective, and ethical over time.