Customer Churn Prediction in Banking: A Machine Learning Approach

Jiyoon Moon
Student ID: 240145156

1 Introduction

Due to the presence of numerous service providers, the market is highly dynamic and fiercely competitive across most industries. This is one of the biggest challenges companies have in the current market environment – understanding how quickly customer behaviors are changing and how best to meet the rising expectations of customers.

In contrast to previous generations, modern consumers place a high premium on connectivity and personalized service. The abundance of information and the high level of education have left consumers quite informed on the purchasing decisions they make. This leads to what is referred to as analysis paralysis where people overanalyze options and make slow decisions. This change in consumer behavior explains why companies must develop new strategies to meet the changing needs of customers and provide them with more value. In the financial sector, customers can easily switch to another company if they get better service or a better price. Companies know that it is expensive and time consuming to attract new customers than to keep the existing ones. But this does not mean that the companies are not facing the challenge of providing good services in good time while maintaining good relations with the customers.

To overcome this issue, businesses must thoroughly understand and address customer needs, with a special focus on customer churn prevention. Customer churn is the term used to describe the situation in which the customer leaves the company or ceases to use its services; it is a model that has a direct relation with profitability. [1]

The goal of this project is to determine the key factors that cause churn and then to choose the most accurate yet interpretable model so that the bank can use it for improving its customer retention strategies. To do this, the four machine learning models used in this project to predict customer churn are Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Decision Tree.

The dataset has 10 features and 1 target variable that describe customers' demographic and financial characteristics. To address customer churn, four machine learning models are developed and compared to determine their effectiveness in predicting churn and identifying critical features.

Table 1: Description of Dataset Features

Feature	**Description**
Credit Score	Credit score of the customer
Country	Country of residence
Gender	Gender of the customer
Age	Age of the customer
Tenure	Years as a bank client
Balance	Account balance of the customer
Products Number	Financial products owned
Credit Card	Whether the customer has a credit card
Active Member	Whether the customer is an active user
Estimated Salary	Estimated salary of the customer
Churn	Target variable indicating customer churn

- **Logistic Regression:** It is a statistical model that is used to predict the probability of customer churn as a function of a linear combination of input variables.
- **K-Nearest Neighbors (KNN):** A non parametric model that classifies customers by their proximity to other training samples, with the idea being that similar situations have similar outcomes.
- **Support Vector Machine (SVM):** A powerful classifier that searches for the optimal hyperplane to distinguish between churners and non churners.
- **Decision Tree:** A rule based model that partitions the data based on feature relevance and generates a decision tree that can help in explaining the cause of a decision.

2 Exploratory Data Analysis (EDA)

To understand the characteristics of the dataset and potential trends, an exploratory data analysis (EDA) was conducted. Visualization techniques including scatter plots, histograms, and correlation heat maps were used to analyze the relationship between features and customer churn. From the EDA, the relevant features were selected and preprocessed to ensure that the models were trained on meaningful data.

- **Pairwise Relationships:** Scatter plot matrix shows some visible patterns between churn and features like age, balance, and the number of products. Churn is more frequent among customers with higher balance and few products.
- **Feature Correlation:** The correlation heatmap shows that most features have weak correlations with each other, meaning no high collinearity.

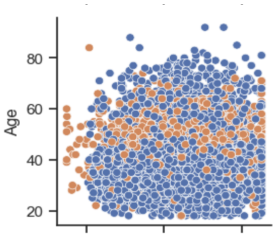

Figure 1: Age vs. Churn

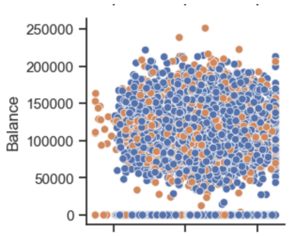

Figure 2: Balance vs. Churn

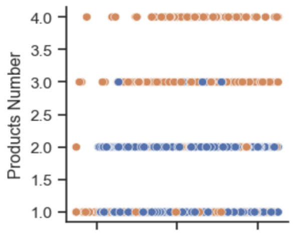

Figure 3: Products vs. Churn

There is a positive correlation between the number of products, balance, and estimated salary. A weak negative correlation of -0.3 exists between churn and the number of products, indicating customers with more products are less likely to churn. Credit score, tenure, and estimated salary have negligible effects on churn.

- **Churn Distribution:** This dataset is heavily imbalanced; a small proportion of customers have churned compared to those who have not. This imbalance needs to be considered to avoid biased predictions.

3 Methods

3.1 Logistic Regression with L2 Regularization

3.1.1 Model Explanation

Logistic regression is one of the most popular techniques for predicting a binary outcome variable such as customer churn using the probability of event occurrence as a function of several input features. L2 regularization, or ridge regularization, is very often used in order to improve the model's generalization ability and avoid overfitting. This adds a penalty term to the loss function proportional to the squared values of the coefficients, which prevents the model from making too much use of any single feature. An optimal trade off between how well the model fits the training data and how simple the model is, is achieved by including L2 regularization, leading to improved predictive performance on new, unseen data. [2]

3.1.2 Key Observations

The model was found to have an accuracy of 81.03% which is very good for the prediction. From the grid search over different values of the regularization parameter λ, it was found that $\lambda = 0.001$ to have the highest cross-validation accuracy of 81.09%, which shows that L2 regularization is effective in improving the generalization capabilities of the model and preventing overfitting.

Other than that, the ROC analysis also supports the robustness of the model

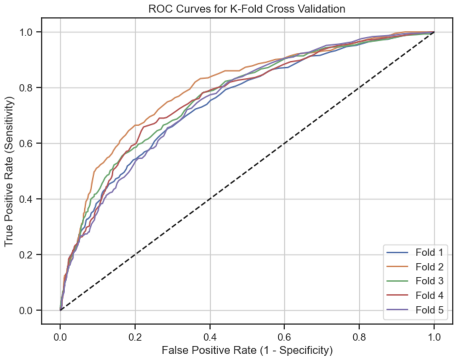

Figure 4: ROC Curves for Logistic Regression

since there is no decline in the performance of the model in any of the five folds of cross-validation. The average accuracy from K-Fold cross-validation was 81.06%, which means that the model is quite stable when the data is divided into different subsets. The shape of the ROC curves indicates that there is a reasonable trade-off between the sensitivity and the specificity, and the AUC values indicate that the models have a moderate discriminative ability.

3.2 K-Nearest Neighbors (KNN) Classification

3.2.1 Model Explanation

The k-Nearest Neighbors (KNN) algorithm is a non-parametric supervised learning method used in classifying new data points by their k closest neighbors. This algorithm finds the k nearest neighbors of the new data point from the training data in the feature space, and it categorizes the new point based on a majority vote. KNN is one of the simplest and most effective classification techniques; however, its efficiency also depends on factors like k value, distance measurement, and complexity in high-dimensional data. [3]

In this project, the KNN model was employed to predict customer churn using the chosen features Age and Balance which were found to be the most critical variables from the exploratory data analysis.

The decision boundary visualization gives a clear view of how well the KNN model classifies the churned and non churned customers. The plot shows that the model is capable of distinguishing between the two classes with reasonable accuracy but there are some regions where the model fails to distinguish between the two classes especially in the area with high density of points.

3.2.2 Key Observations

The accuracy of 75.84% makes it apparent that the KNN model is quite good at predicting customer churn, but it is not as efficient as simpler models, such as

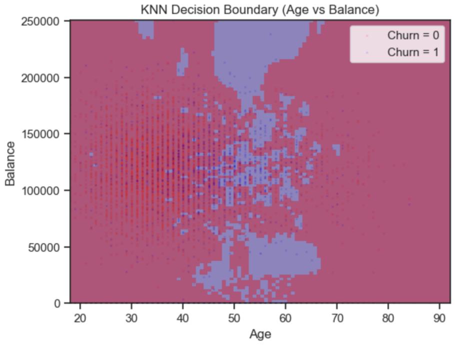

Figure 5: Decision Boundary

logistic regression or decision trees. The choice of $k = 8$ shows that in order to smooth over noise and make better predictions, it needs a larger neighborhood; however, this comes at the cost of reduced sensitivity to the local structure in the data.

The decision boundary plot, which shows where churned and non-churned customers are classified, illustrates that the model has difficulty distinguishing between the two classes, especially in complex, high density regions. In addition, the KNN algorithm is a bad choice for large datasets because it entails calculating distances for every new prediction, and this makes it computationally expensive. Therefore, the results indicate that although KNN is easy to interpret and use, it may not be the best choice for high-dimensional or large datasets because of its inherent computational difficulties. [3]

3.3 Support Vector Machine (SVM)

3.3.1 Model Explanation

The SVM (Support Vector Machines) is a classification algorithm that finds the optimal hyperplane that separates different classes with the maximum margin. This approach transforms the features from the input space to a higher dimensional space so that a linear decision boundary can be applied. These are called support vectors and they help in defining this boundary of the model to be able to learn the complex relationships in the data efficiently. The kernel functions used are linear, polynomial and radial basis function (RBF) to map the input data into a higher dimensional space to classify nonlinearly separable data. The optimization process is to maximize the generalization performance by enforcing a margin while ensuring accuracy of classification. [4]

In this project, the SVM model was applied to identify customer churn to find the optimal hyperplane which distinguishes between the churned and not churned customers. The optimal value of the regularization parameter C was obtained from grid search with 0.1 and the final accuracy of 80.56%.

3.3.2 Key Observations

The accuracy of the SVM model is 80.56% which means it can effectively predict the result for the new data. However, the accuracy is a bit lower than that of logistic regression and decision trees. For $C = 0.1$ we get the best accuracy which implies that smaller values of C are better for maximizing the margin and for generalization. This means that low values of C (i.e., 0.01, 0.1) provide better accuracy than high values of C (i.e., 1, 10), which may suffer from overfitting, where the model is too complex and fits the training data closely.

A grid search over various values of C showed that this dataset performs better with smaller C to avoid overfitting and to keep decision boundary flexible. The selected parameter C gives the right trade off between margin maximization and misclassification penalties and hence improves the generalization.

The SVM model is formulated as a quadratic optimization problem and can become computationally expensive for large dataset. Even though SVM works well for moderate sized datasets, it may not be suitable for very large datasets without using kernel approximations or feature reduction. [4]

3.4 Decision Tree with SMOTE Oversampling

3.4.1 Model Explanation

The decision tree is a classification algorithm that forms data into a decision tree which classifies data on the basis of features. It divides the feature space into a number of distinct, non-overlapping regions and places a class label at the bottom of each region. The tree is constructed using measures such as Gini impurity or entropy to determine the optimal split at each level. They are easy to interpret and are even efficient even for large datasets; however, they tend to overfit which can be reduced through pruning. [5]

Metric	**Before SMOTE**	**After SMOTE**
Accuracy	85.89%	87.16%
Precision	79.25%	85.35%
Recall	41.63%	74.94%
F1 Score	54.59%	79.81%

Figure 6: Comparison of Model Performance

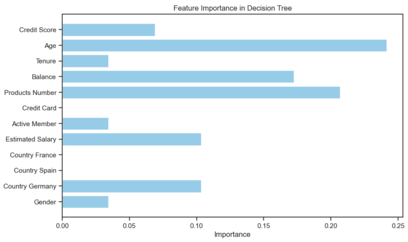

Figure 7: Feature Importance

3.4.2 Key Observations

A decision tree model was used in this project to predict customer churn. Cross validation accuracy of 85.89% was achieved with a tree depth of 5 when the model was trained using grid search. Age, products number and balance were revealed to be the most important features in predicting churn through feature importance analysis. The model was very good at identifying which customers did not churn but had difficulty with the customers who did churn, with a low recall of 41.63%.

The dataset was unbalanced at first because it had a lot of customers who did not churn and very few customers who did. This unrecoverable bias caused the model to focus too much on not identifying churned customers when it failed to identify how to identify them properly. To relieve this problem, the Synthetic Minority Over-sampling Technique (SMOTE) was used to create new samples of the minority class (churned customers). This made the dataset more balanced to make the model to be able to learn the patterns of the churned customers.

The algorithm identifies the k nearest neighbours of a minority class point and creates synthetic instances on the path between the point and its neighbours. This improves the contribution of the minority class to the decision boundary, leading to improved model generalisation and reduced overfitting. [5]

Applying SMOTE has effectively improved the decision tree model's ability to identify churned customers by equitably handling the class imbalance problem. The enhancement in recall from 41.63% to 74.94% and the improvement in F1 score from 54.59% to 79.81% are particularly important for customer retention programs because it enables the bank to prevent the customer defection from happening.

The increased recall rate means that more actual churned customers are likely to be identified correctly, while the improved F1 score suggests that the tradeoff between precision and recall is better, so that there are fewer false positives and negatives. This enables the bank to design more specific and efficient campaigns for customer retention, thus enhancing customer satisfaction and lessening the financial impact. [1]

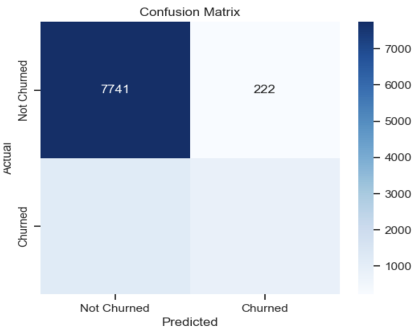

Figure 8: Confusion Matrix Before SMOTE

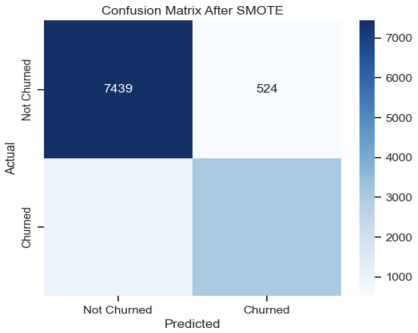

Figure 9: Confusion Matrix After SMOTE

4 Conclusion

This project employed different machine learning techniques to develop a model that can help the bank to identify customers who are likely to leave based on demographic and financial characteristics of bank customers. The goal was to identify potential causes of customer churn and to identify the most appropriate model to help the bank to design effective customer retention plans. Different machine learning models were developed and compared to see which of the models performed better in predicting customers' churn.

The accuracy of logistic regression with L2 regularization was 81.00%, and this accuracy did not vary significantly across different data splits in cross-validation. The KNN model had a slightly lower accuracy of 75.84%, and it was very good at categorizing the customers as churned or not but was not efficient with close data points. The SVM model gave an accuracy of 80.56%, and the best value of the regularization parameter C led to good results. However, SVM is a computationally expensive method and is sensitive to noise. The decision tree model initially had the highest accuracy of 85.89%, but it had a very low recall of 41.63%, which means it was bad at identifying the real churned customers, although it was good overall.

To reduce the class imbalance problem and to improve the ability of the model to identify the churned customers, SMOTE (Synthetic Minority Over-sampling Technique) was employed. After applying the decision tree model with over-sampling, the accuracy increased to 87.16%, recall to 74.94%, and F1-score to 79.81%.

Although, there are some implications for further development. Some of them are: Using other methods such as Random Forest and Gradient Boosting is worth a try since they might improve the accuracy and the ability of the model to generalize. It would also be helpful to incorporate new features that are derived from customer transactional data and interaction to get a better understanding of the churn. Since there is the issue of precision versus recall, it may be helpful to look at cost-sensitive learning to find the optimal model. Finally, applying the model in real life and updating it periodically will help it to adapt to changing environment of customers. [6]

In conclusion, the decision tree model with SMOTE oversampling was identified as the most accurate yet easy to interpret model for determining customer churn. As such, the enhanced capacity to pinpoint churned customers through oversampling is particularly useful for customer retention management activities, thereby giving the bank the opportunity to act before customers defect. However, it is still important to keep on perfecting and improving the model so that it can remain applicable and flexible with regard to shifting customers' behaviors.

References

[1] M. Rahman and K. Vasimalla, "Machine learning based customer churn prediction in banking," in *Proceedings of the International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, November 2020, p. 9297529.

[2] A. Y. Ng, "Feature selection, l1 vs. l2 regularization, and rotational invariance," *Proceedings of the 21st International Conference on Machine Learning*, vol. 21, pp. 78–86, 2004.

[3] P. Cunningham and S. J. Delany, *k-Nearest Neighbour Classifiers (2nd Edition with Python examples)*. Technological University Dublin, 2020. [Online]. Available: https://github.com/PadraigC/kNNTutorial

[4] J. M. Moguerza and A. Muñoz, "Support vector machines with applications," *Statistical Science*, vol. 21, no. 3, pp. 322–336, 2006.

[5] H. Chawla, Bowyer and Kegelmeyer, "Smote: Synthetic minority oversampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[6] J. Maan and H. Maan, "Customer churn prediction model using explainable machine learning," *International Journal of Computer Science Trends and Technology (IJCST)*, vol. 11, no. 1, pp. 33–38, 2023. [Online]. Available: http://www.ijcstjournal.org/