Predicting Customer Churn Using Decision Trees

Abstract

This project applies the Decision Tree algorithm to predict customer churn using a public banking dataset. The model was trained on customer demographic and account features and fine-tuned using cross-validation and hyperparameter optimization. The final model achieved strong predictive performance, with an overall F1-score of 0.82 and ROC AUC of 0.84. This solution enables banks to identify potential churners and take proactive steps to retain them.

Problem Statement

Banks face a continuous challenge in retaining their customers, especially in a competitive market where switching costs are low. Predicting customer churn - identifying clients likely to leave the bank - is critical for implementing timely retention strategies. This project develops a Decision Tree model to classify whether a customer will churn based on their demographic and financial attributes. The model is optimized to balance precision and recall while maintaining interpretability.

1. Dataset Overview

- Source: Kaggle (Bank Customer Churn Dataset)
- Target Variable: churn (0 = stayed, 1 = left)
- Key features: credit score, age, gender, country, balance, tenure, etc.
- Dataset Size: ~10,000 entries

2. Data Preparation & Preprocessing

To ensure the data was suitable for modeling:
- Irrelevant feature `customer_id` was dropped
- Dummy encoding was applied to categorical features like `country` and `gender`
- Synthetic Minority Oversampling Technique (SMOTE) was used to balance the class distribution
- Data was split into training and test sets using a 75/25 ratio

3. Model Building

Decision Tree Classifier
Two models were trained:
- Base Model: A default Decision Tree using Gini impurity
- Fine-Tuned Model: Optimized using GridSearchCV with 5-fold cross-validation
  - Criterion: Gini
  - Max Depth: 10
  - Min Samples Split: 10
  - Min Samples Leaf: 2

4. Model Evaluation & Comparison

The performance of both models was evaluated using Accuracy, F1-score, and ROC AUC. Visual diagnostics including ROC curves and boxplots of predicted probabilities were also used.

Model	Accuracy	F1 Score	ROC AUC
Base Model	77.0%	0.77	0.73
Fine-Tuned	82.0%	0.82	0.84

5. Discussion & Insights

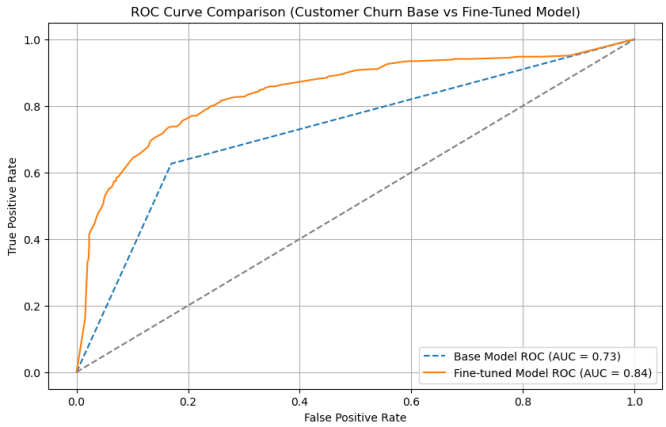

Figure 5.1: ROC curves comparison

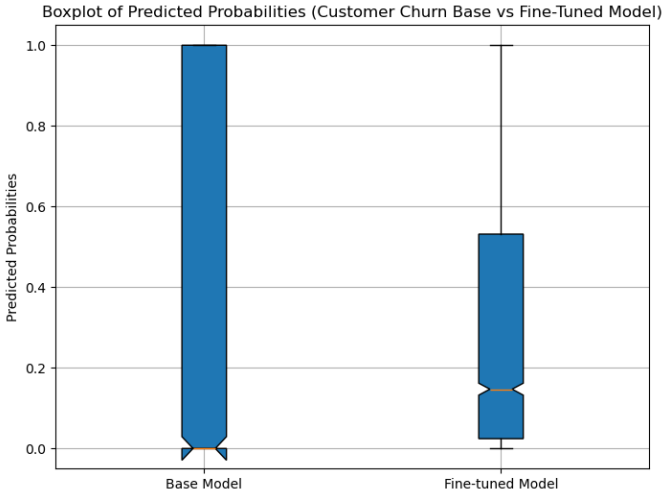

Figure 5.2: *Boxplot of Predicted Probabilities (Base vs Fine-Tuned Model)*

The fine-tuned Decision Tree model significantly outperformed the base model across all metrics. Most notably, the overall F1-score improved to 0.82, indicating strong generalisation and balanced prediction performance. The use of SMOTE helped the model

better identify churners, while hyperparameter tuning reduced overfitting and improved stability. These improvements are critical in practical banking applications where even small gains in prediction can lead to better customer retention.

6. Conclusion

This project demonstrates that a well-tuned Decision Tree classifier can effectively predict customer churn in the banking sector. Its interpretability and solid performance make it suitable for stakeholder-facing applications. Future work could involve ensemble models such as Random Forest or XGBoost to further enhance accuracy and robustness.