Banking Customers Churn Prediction - Final Delivery

João Mendes Silva Belchior - up202108777
José Francisco Reis Pedreiras Neves Veiga - up202108753
Pedro Vidal Marcelino - up202108754

Problem definition

- Accurately predict if any customers are planning on exiting a bank, using the given dataset.
- The dataset has 10000 entries, each with the following columns:

  CustomerID | Surname | Credit Score | Geography | Gender | Age | Tenure | Number of products | Credit Card Ownership | Active | Salary | Exited

Problem classification

- Supervised Learning
- Binary Classification Problem

Related work

1. For the same dataset we've found 6 solutions developed by others with accuracies rounding 70 -80 % for most algorithms https://www.kaggle.com/datasets/saurabhbadole/bank-customer-churn-prediction-dataset/code
2. Machine Learning in Healthcare Projects: From Not-For-Everyone Treatment To Mass-market https://www.aimprosoft.com/blog/machine-learning-in-healthcare/
3. Machine learning techniques for classifying dangerous asteroids https://www.sciencedirect.com/science/article/pii/S2215016123003345

Tools

- Pandas
- NumPy
- Scikit-learn
- Seaborn
- Matplotlib

Algorithms

- Decision Tree based Methods
- K-Nearest-Neighbor
- Naïve Bayes
- Support Vector Machines
- Neural Networks, Deep Neural Network

Implemented Work

Data preparation:

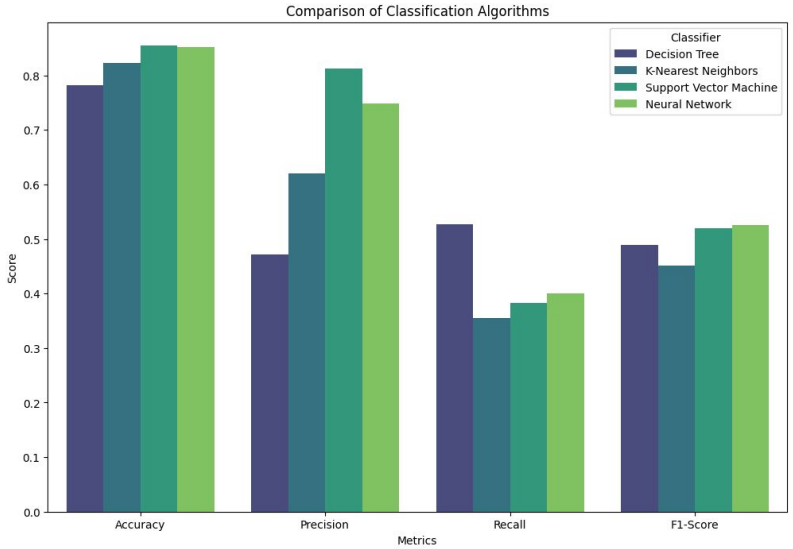
- Drop unnecessary columns (RowNumber, CustomerId and Surname)
- Check for erroneous values (null, negative, missing, unusual)
- Handle categorical variables (Geography: one-hot encoding, Gender: binary encoding)
- Need to handle imbalanced churned vs retained

Modeling:

- Decision Tree, K-NN, SVM, Neural Network
- Accuracy (Overall correctness; proportion of correctly classified instances)
- Precision (Ability to avoid false alarms; true positives among predicted positives)
- Recall (Ability to capture positive instances; true positives among actual positives)
- F1-score (Balanced measure of precision and recall; harmonic mean of precision and recall)

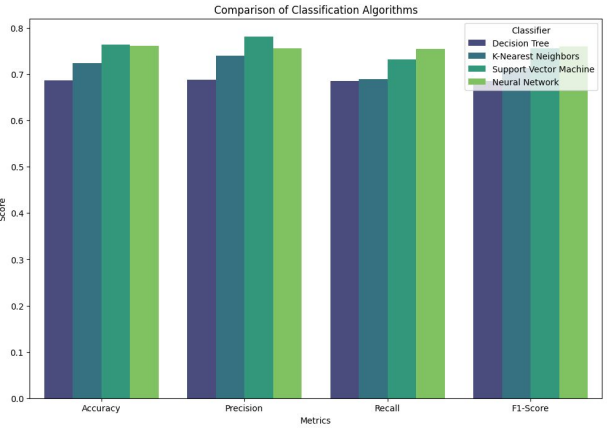
Model Improvement

Initial Results

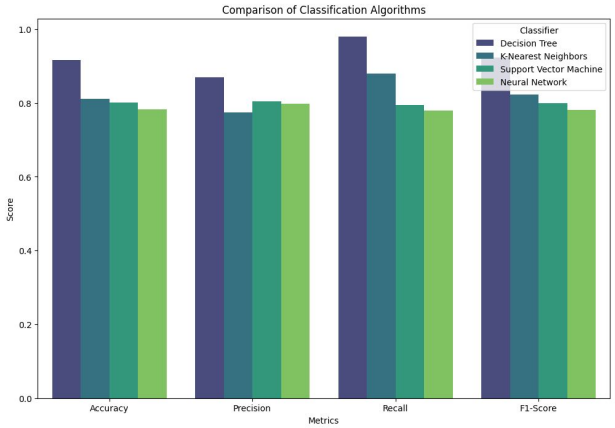


- Balance the Dataset using techniques such as Random Over Sampling, SMOTE, and Random Under Sampling to address class imbalance.
- Use hyperparameter tuning with Grid Search.

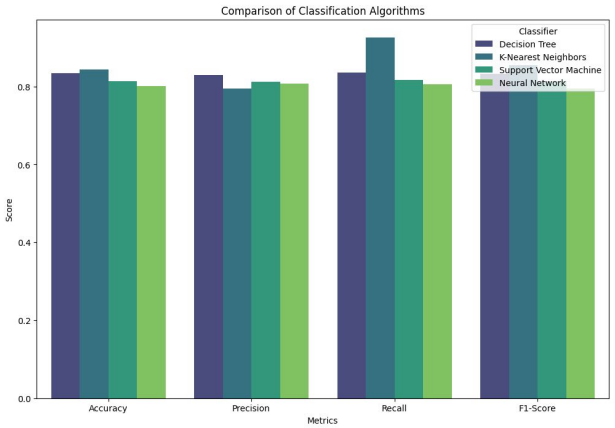
Model Improvement (Balanced only)



Under-sample

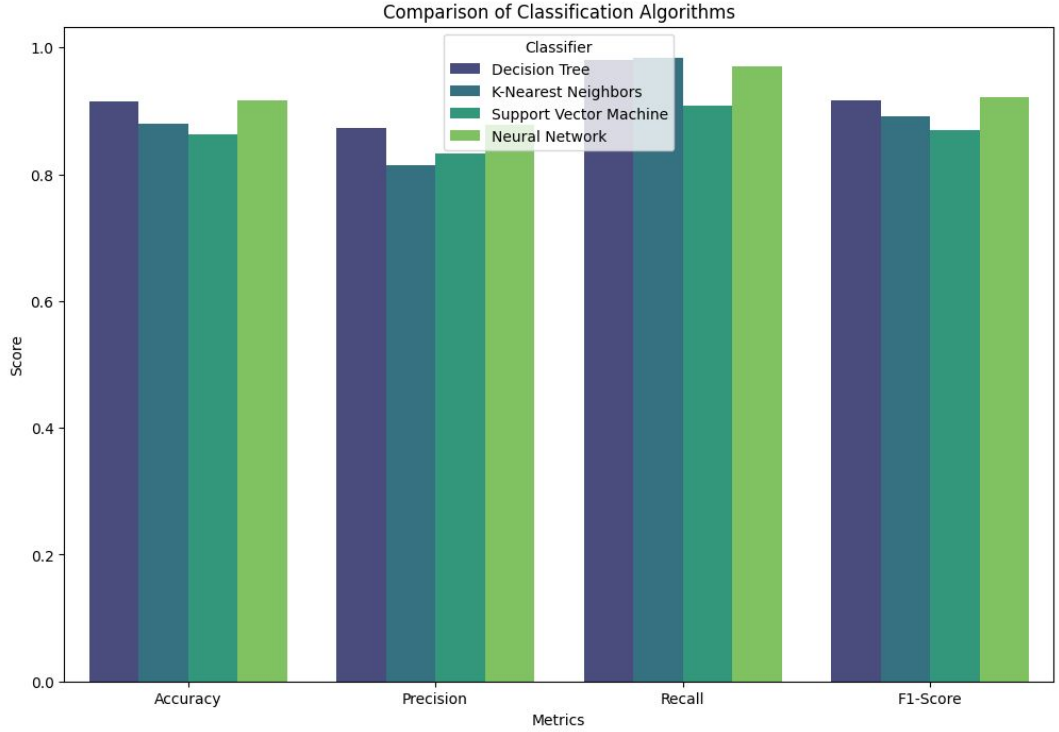


Over-sample



Smote

Model Improvement (Hyperparameter tuned)



An over-sampled dataset is the only one that saw improvements from this tuning, due to:

- An increase in the minority class representation
- Not losing any data from the majority class
- Leveraging more data during training

Conclusions

The best model for this churn prediction problem was identified and optimized.

Balancing the dataset significantly improved model performance.

Hyperparameter tuning only showed improvements when using the over-sampled dataset.

Deploying this model will enable the bank to proactively retain customers, reducing churn and increasing customer satisfaction.