STRATEGIC THINKING

Banking Industry

Ariel Goldman

BANK CHURNERS

=

Business Understanding

Our project is focused on the banking industry, with the main goal of improving and retaining the customers to reduce churn rates.

Hypothesis

Analyse why customers will churn or not from the banks based on their demographic features and transaction history.

General Goal

Develop a predictive Machine Learning Model which can accurately classify customers as either churn or not.

Success Criteria/Indicators

The success of our proyect we will be measured by the accuracy of our Machine Learning Models in predicting customer churn.

Technology Used

Machine Learning Models

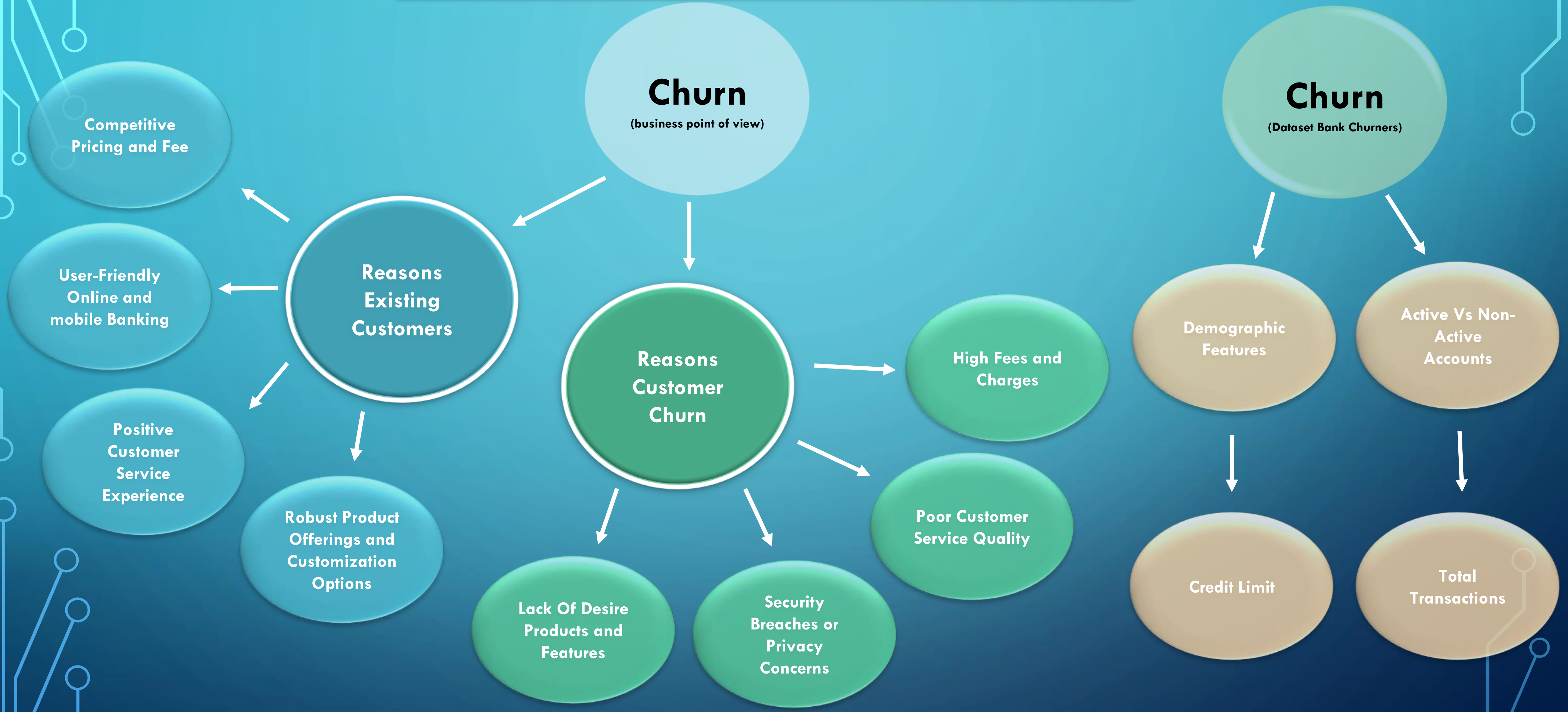
- Decision Tree
- SupportVector Machine
- Logistic Regression
- AdaBoost
- Random Forest
- Gaussian Naïve Bayes
- KNeighbours

Libraries

- Scikit-learn
- Seaborn
- Matplotlib
- Pandas
- NumPy
- Imbelear.over_sampling.SMOTE
- Category_encoders
- Missigno

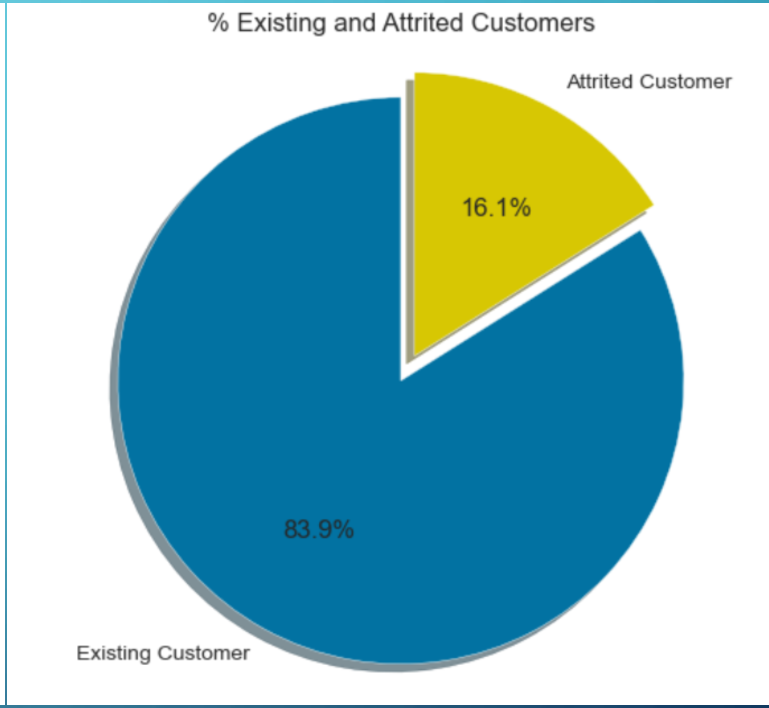
Factors Affecting Customer Churn and Retention in Banks.

Conceptual Understanding of the business

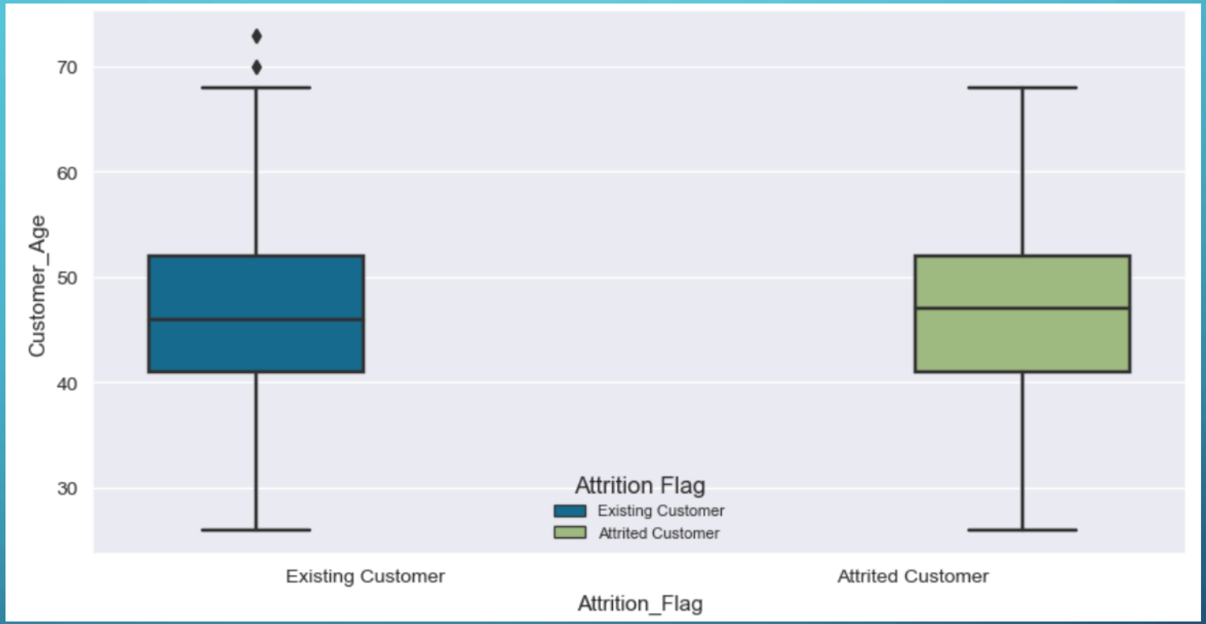


Exploratory Data Analysis

Imbalaced Dataset

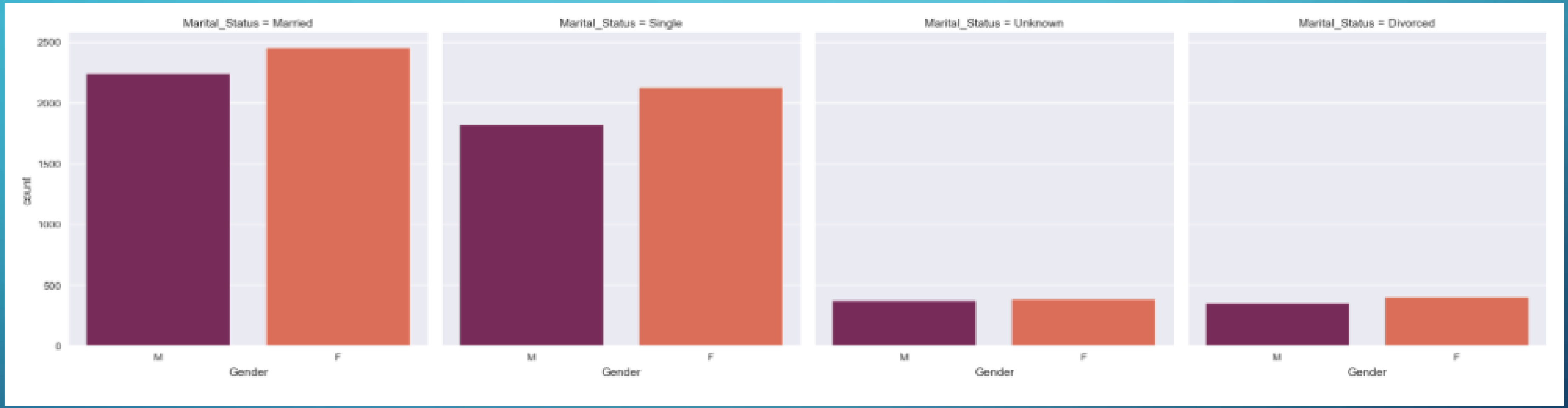


It seems that older customers are more likely to leave the bank.

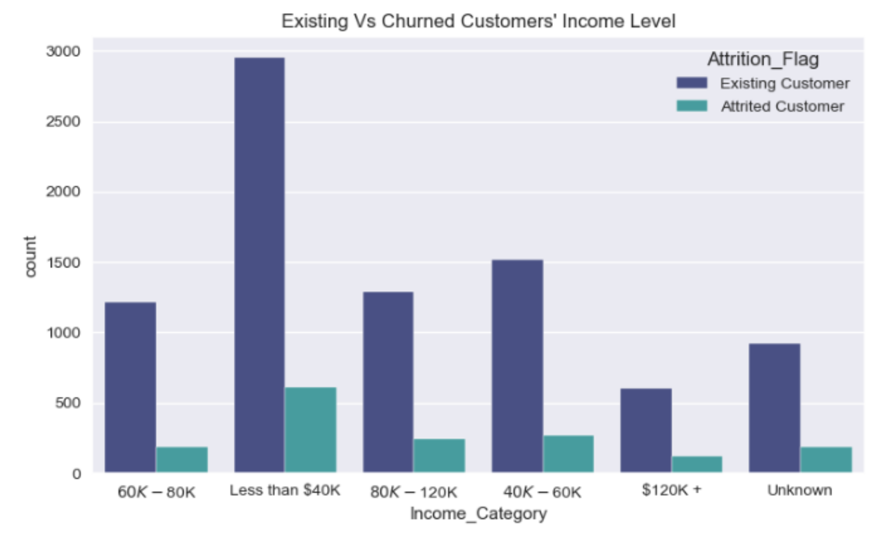


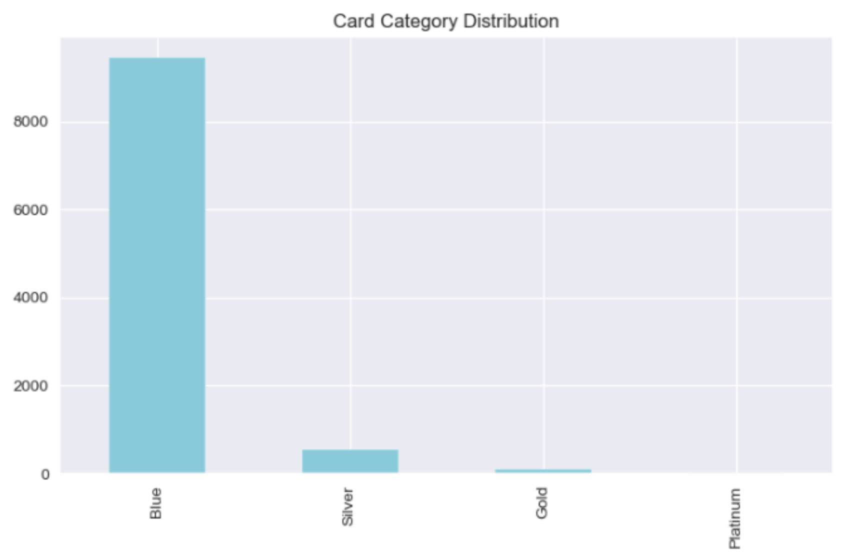
Customers married and single are the majority in the bank.

- Special promotions
- Incentives



- The mayority of the customers have an income less than $40k.
- 93% of the customers have credit card ‘Blue’.





Exisitng Customers

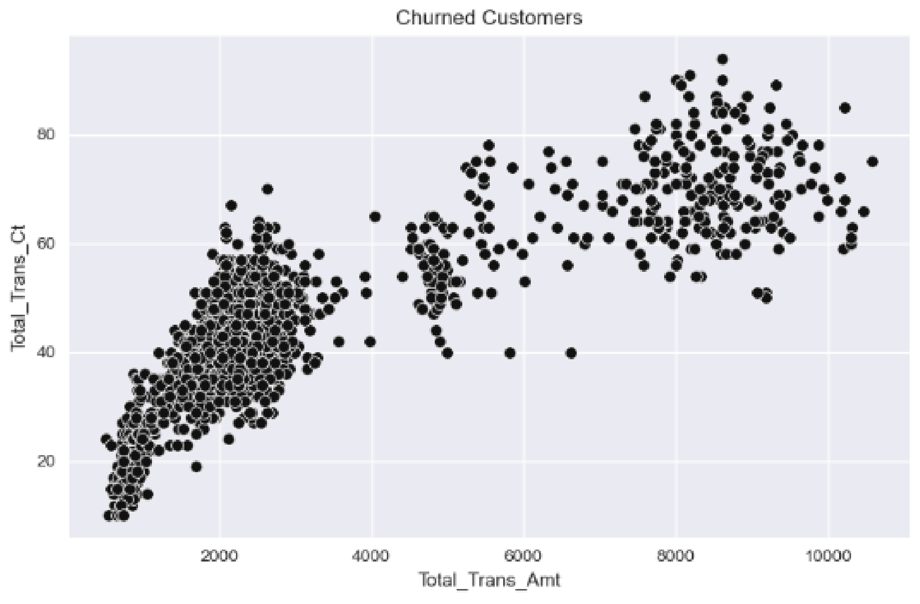
Total Transaction counts Vs Total Transaction Amount

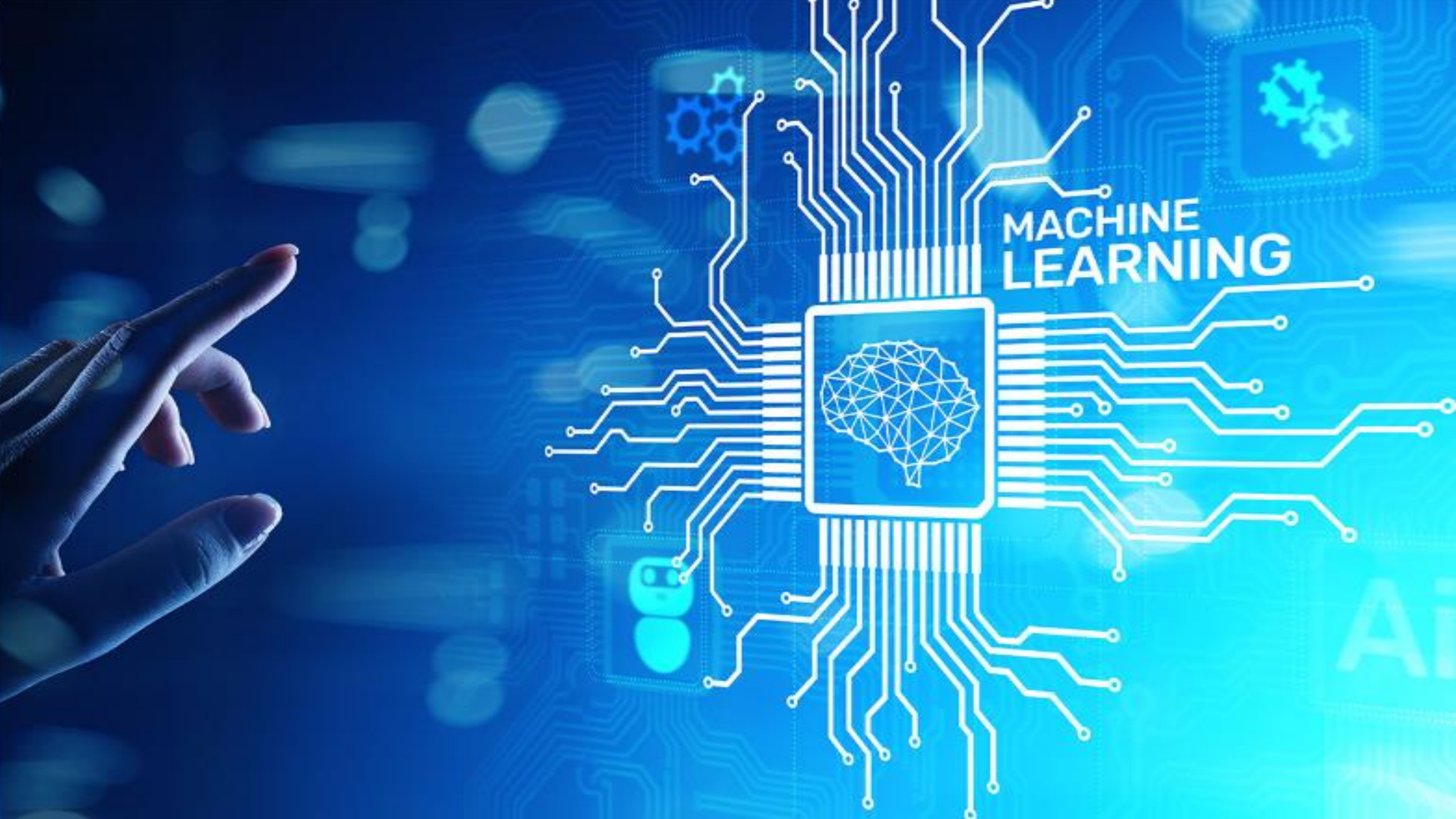
3 distinct clusters

Churned Customers

Total Transaction Counts Vs Total Transaction Amount

1 distinct cluster




MACHINE
LEARNING

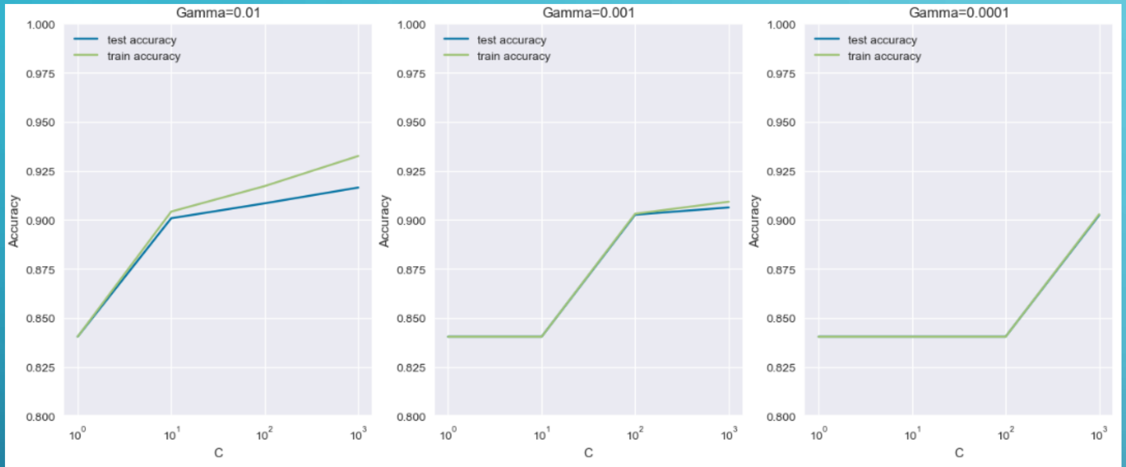
Code Development

- Finding Suitable Dataset
- Data Processing
- Exploratory Data Analysis
- Feature Engineering
- Select Appropriate ML Model
- Training The Model
- Evaluate Performance
- Model Tuning
- Deployment

Deployment of Machine Learning Model
SVM and Random Forest

SVM

- Supervised learning
- Classify customers as churn or non-churn based on banking behavior
- GridSearchCV to find optimal hyperparameter ('Kernel=RBF')



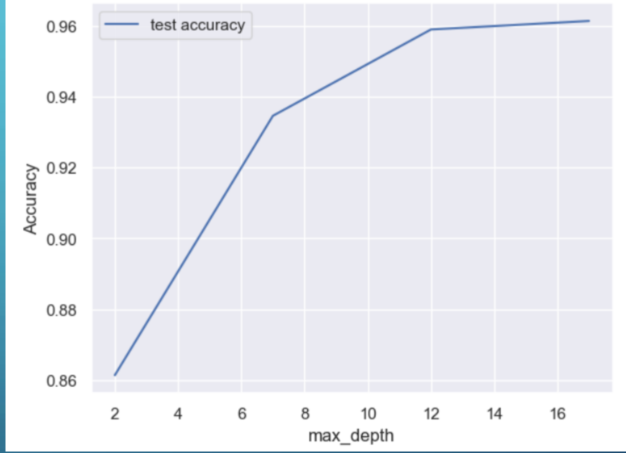
RBF Linear Kernel Model

- The model has a good balance between being too simple (underfitting) and being too complex (overfitting).
- The best test score is 92%, it means the model is able to predict the correct output (generated by the model) based on the provided data.

Random Forest

- Supervised learning
- Classify customers as churn or non-churn based on banking behavior
- Hyperparameter Tuning - Max_depth

params	split0_test_score	split1_test_score	split2_test_score	split3_test_score	split4_test_score	mean_test_score	std_test_score	rank_test_score
{'max_depth': 2}	0.86	0.87	0.88	0.87	0.87	0.87	0.00	4
{'max_depth': 7}	0.94	0.93	0.94	0.94	0.94	0.94	0.00	3
{'max_depth': 12}	0.96	0.96	0.96	0.95	0.96	0.96	0.00	2
{'max_depth': 17}	0.96	0.96	0.96	0.96	0.97	0.96	0.00	1

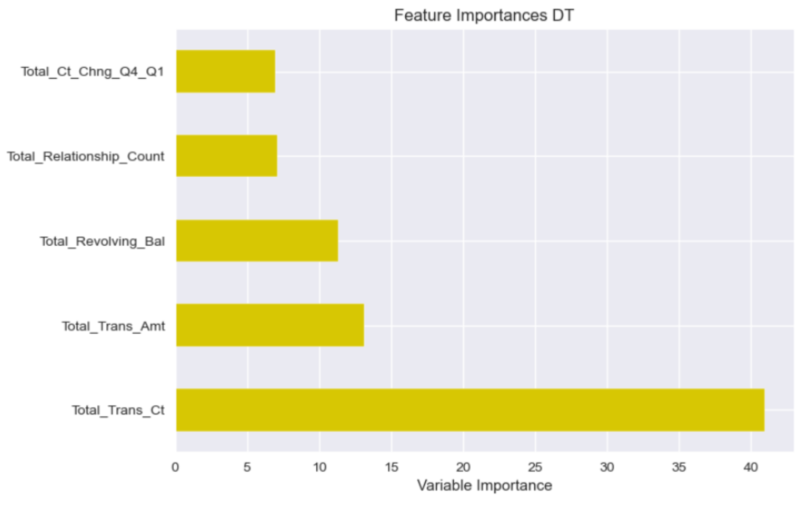


After fitting the model with the values of 'max_depth', the plot indicates that the test accuracy 96% is high when 'max_depth' is set in a range of 2 to 20 while splitting the model into 5.

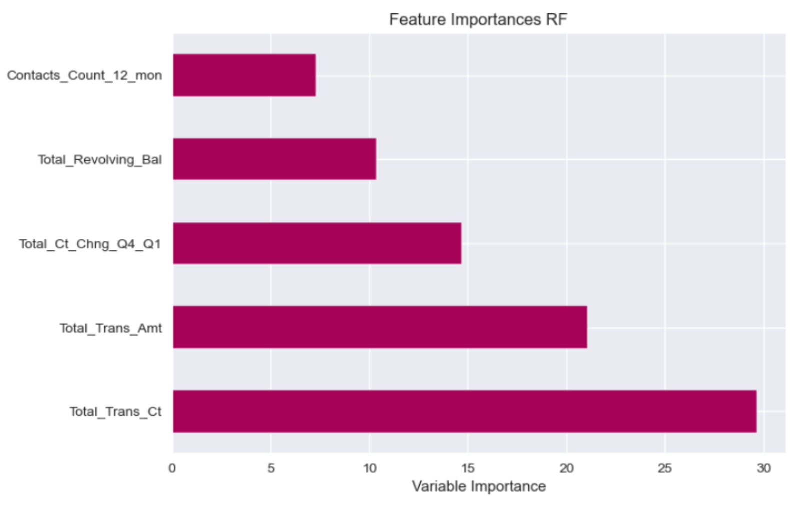
Future Importances

Total Transaction Count stands out across the three analysed models.

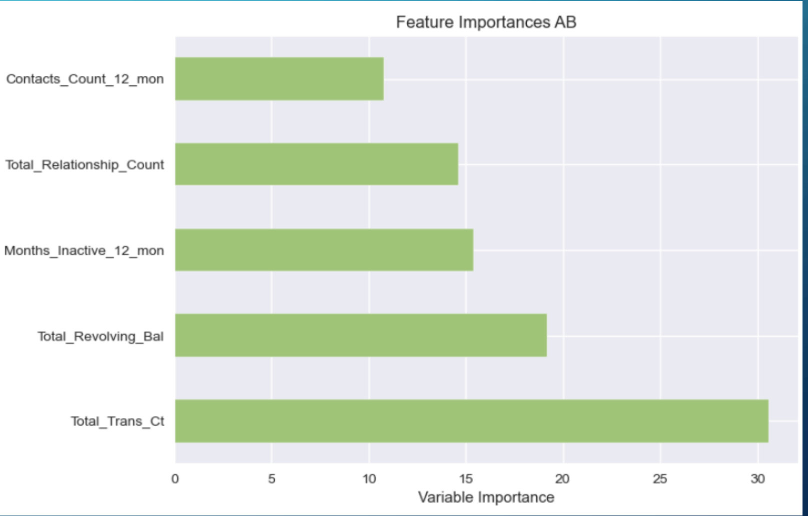
Decision Tree



Random Forest

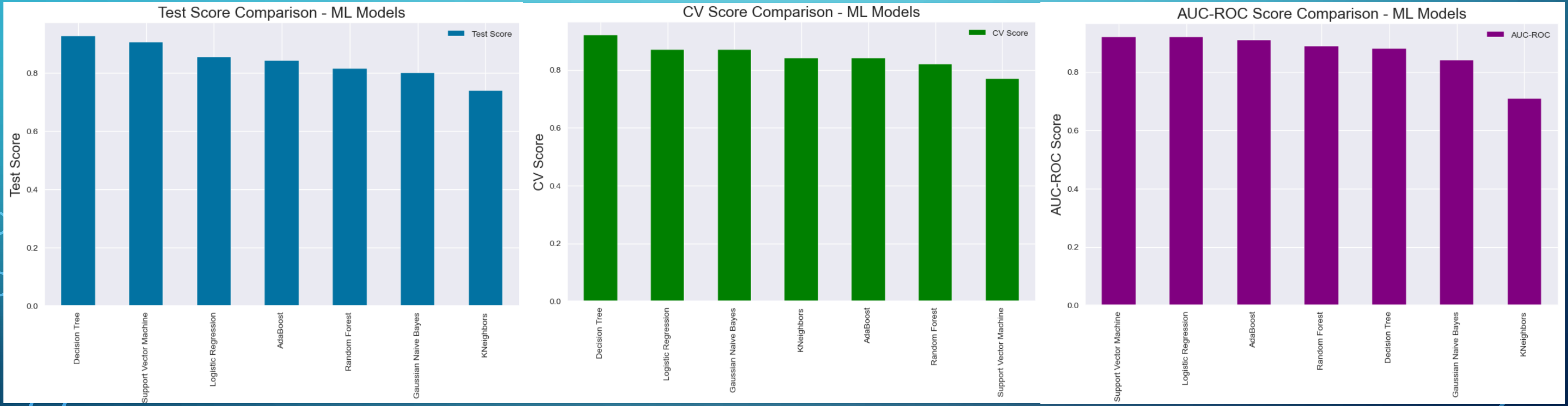


AdaBoost



Model Comparisons

After evaluating the performance of the seven trained models, it appears that Support Vector Machine and Logistic Regression models are consistently performing well across all models. They achieved high scores in both tests scores and AUC-ROC. Decision Tree is also performing well when the model was cross validated.



Conclusion

Demographic factors like age, gender, income, marital status, and education are not reliable indicators for predicting customer churn, as shown by Random Forest, Decision Tree, and Adaboost models. Even after an extensive Grid Search, no clear correlation was established. On the other hand, transaction history features consistently proved highly significant. In conclusion, further investigation is needed to identify the most effective model for real-world deployment.

QUESTIONS?

Thank you !!!