Predictive Analytics for Customer Churns in Financial Services Industry

Author: Thant Thiha | Supervisor: Taufique Ahmed

cct | College Dublin Computing • IT • Business

Abstract

This project addresses the critical challenge of customer attrition in banking. By leveraging machine learning, we achieved an **86% ROC-AUC in predicting churn**. The solution identifies high-risk customers and key drivers (such as product usage and age), enabling targeted, cost-effective retention strategies.

Business Problems

The Cost of Churn: Acquiring a new customer costs 5-7 times more than retaining an existing one.

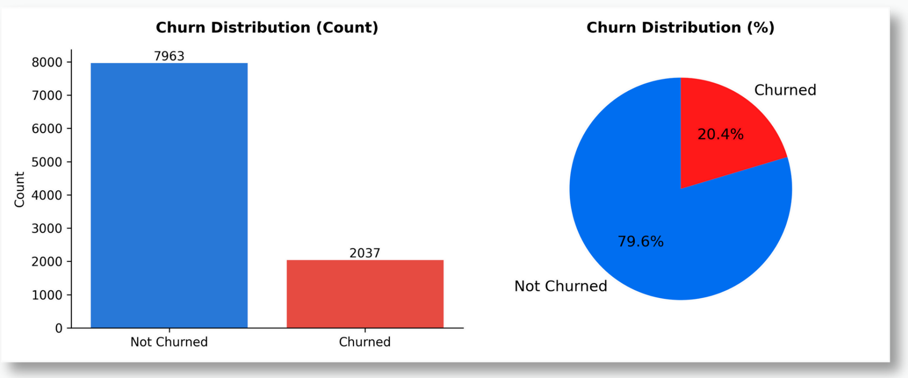
Objective: Develop a predictive model to identify at-risk customers before they exit.

Success Criteria:

- ROC-AUC Score $> 75\%$.
- Identification of Top 5 Churn Drivers.
- Explainability for customer service intervention.

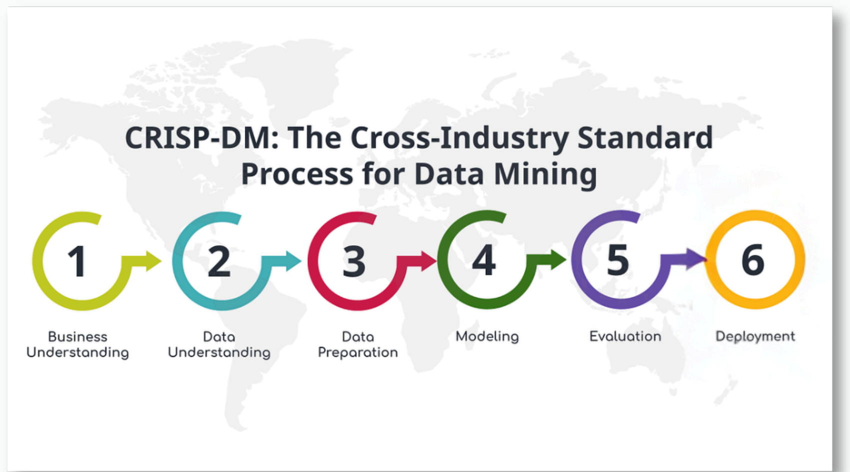
Dataset

The dataset consists of 10,000 anonymized customer records with 14 features covering demographic, behavioral, and financial attributes with **20.4% churn rate** creating a 4:1 imbalance that required synthetic oversampling.



Methodology: CRISP-DM Framework

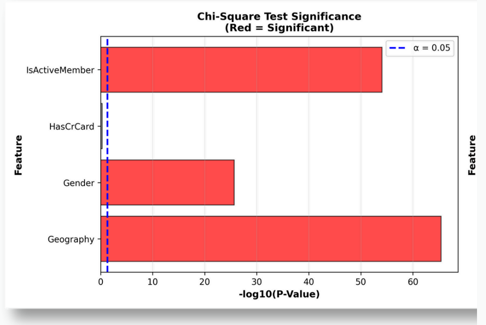
1. **Business Understanding** - Defined retention objectives and KPIs
2. **Data Understanding** - EDA + statistical hypothesis testing (Chi-Square, Mann-Whitney U)
3. **Data Preparation** - Feature engineering, SMOTE balancing, StandardScaler, 80/20 split
4. **Modeling** - 5 algorithms, GridSearchCV, 5-fold CV
5. **Evaluation** - Multi-metric comparison, SHAP/LIME interpretability, bias detection
6. **Deployment** - technically ready for deployment

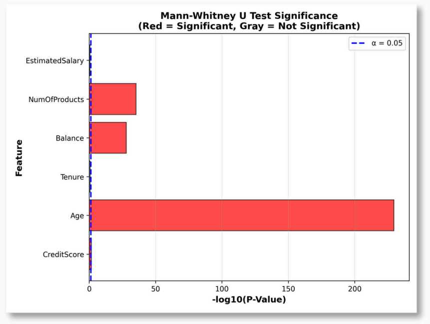


Data Understanding and EDA

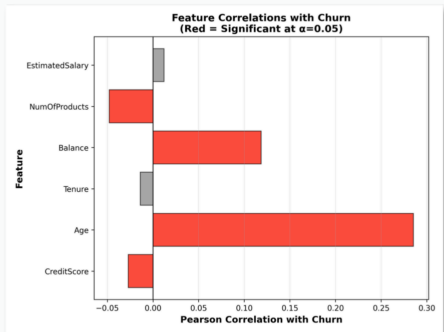
Distributions and Statistical Insights

- **Average Age of 38.9** years old and **Average Tenure of 5 years**
- **Geography**: *Germany* shows *highest churn rate (32.4%)* vs France/Spain (~16%).
- **Activity**: *Inactive members churn at 26.9%* vs Active (14.3%) (validated via Chi-square Test).
- **Age**: Churners are statistically *older* (validated via Mann-Whitney U test).





- Age, Balance, Number of Products, and Credit Scores are weakly correlated with Churn.



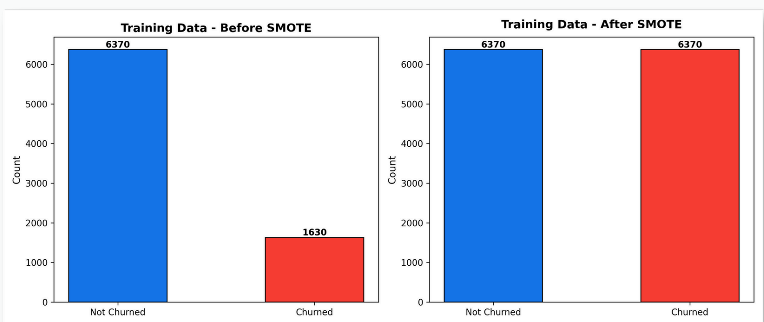
Data Preparation

Feature Engineering To capture non-linear risks, we engineered five specific features:

- **Age_Group** & **Is_Senior**: To isolate the higher risk observed in older demographics.
- **BalanceToSalary_Ratio**: To gauge financial commitment relative to income.
- **Has_Zero_Balance**: To identify dormant accounts.

Preprocessing and Strategy

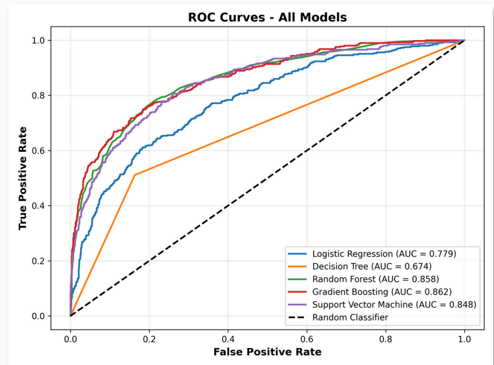
- **Encoding**: ***One-Hot Encoding*** was used for **Geography** to prevent ordinal bias; ***Binary*** encoding for Gender.
- **Scaling**: Numerical features (Balance, Estimated Salary) were normalized using ***StandardScaler*** to ensure model convergence.
- **Handling Imbalance** (Crucial Step): We applied ***SMOTE (Synthetic Minority Over-sampling Technique)*** to the training set. This balanced the classes 50/50, ensuring the model learned to detect churners rather than just guessing "Retained" for everyone.



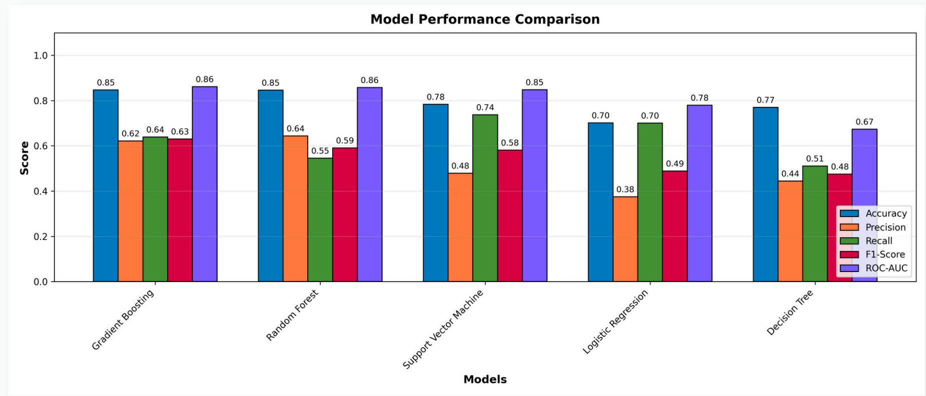
Modeling & Evaluations

We benchmarked five algorithms: **Logistic Regression, Decision Tree, SVM, Random Forest, and Gradient Boosting.**

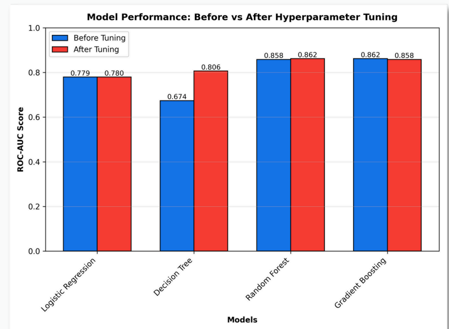
- **Winner**: Ensemble models (***Random Forest and Gradient Boosting***) outperformed single learners.
- **ROC-AUC**: ~86% (Exceeded 75% target).
- **Accuracy**: 85%
- **Reliability**: Confirmed via 5-Fold Cross-Validation with low standard deviation.



- **Comparison**: Logistic Regression lagged at ~78% ROC-AUC, failing to capture non-linear patterns.



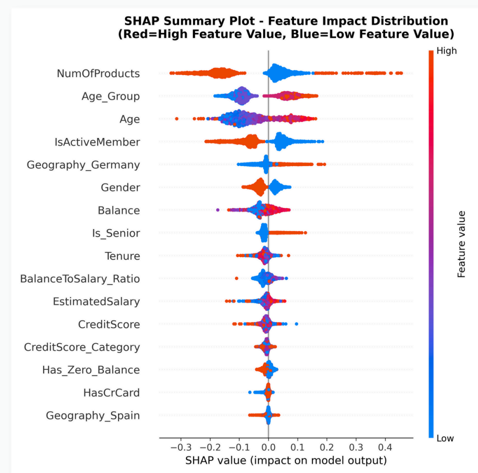
Optimization: Hyperparameter tuning (GridSearchCV) significantly boosted the Decision Tree (ROC-AUC +0.14), while fine-tuning the Random Forest for maximum stability.



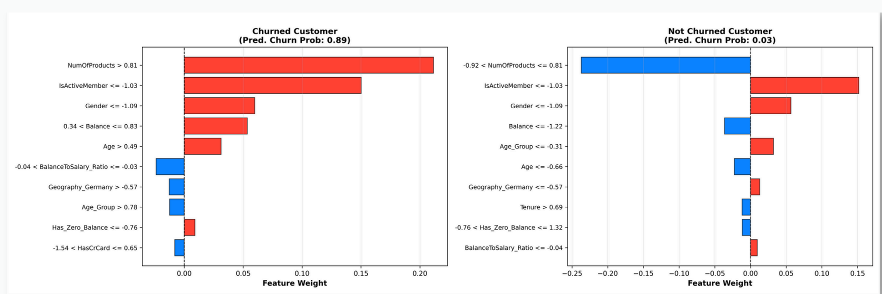
Drivers of Churns

Explainable AI (XAI) To ensure the "***Black Box***" models were usable by business teams, we used ***SHAP and LIME***.

- **Top Predictors:**
  a. **Number of Products**: ***Customers with 3+ products*** are highly likely to churn (possible overload/dissatisfaction).
  b. **Age**: Risk rises sharply ***after age 40.***
  c. **Geography**: Customers from ***Germany*** significantly increase risk.



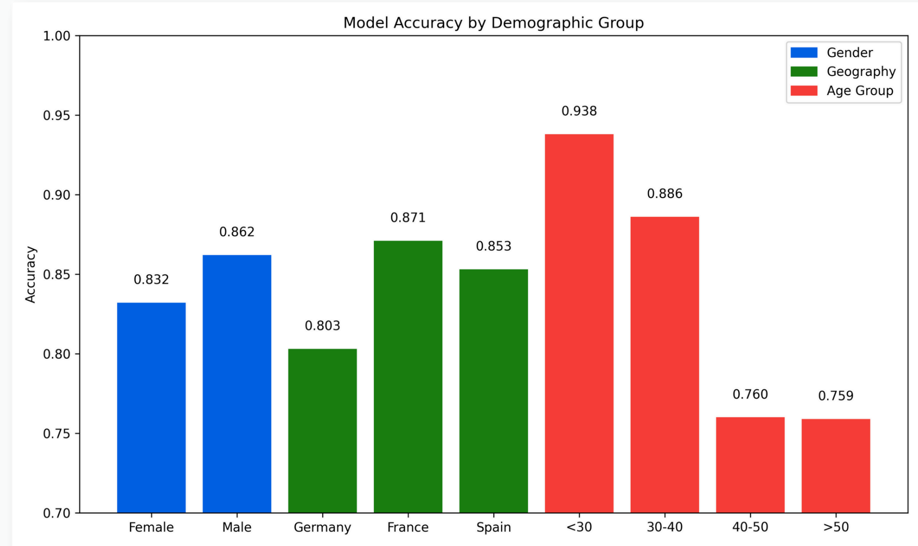
- **Surprise Finding**: ***EstimatedSalary*** and ***CreditScore*** were weak predictors, suggesting churn is driven by service experience, ***not customer wealth.***



Ethical Consideration

We conducted a bias audit to ensure equitable performance and fairness in AI.

- **Findings**: The model is slightly less accurate for ***Female customers (83%)*** and ***German residents (80%).***
- **Mitigation**: Strict monitoring is required for these subgroups. We should implement ***"Human-in-the-loop"*** protocols where high-risk predictions for these groups are reviewed manually.



Business Recommendations

- **Retention Strategy**: Focus budget on Multi-product users and the 40-50 age group.
- **Geographic Focus**: Investigate the German market product offering—churn is abnormally high there.
- **Proactive Outreach**: Target "Inactive" members with engagement campaigns before their tenure reaches the critical 3-4 year mark.

Conclusion

- **Project Status**: Achieved 86% ROC-AUC, validated statistically, and ethically audited.
- **Recommendation**: Deploy the tuned Random Forest model due to its balance of high accuracy and stability.
- **Future Work:** Incorporate real-time transaction data to improve "IsActive" definition and address the accuracy gap in the German demographic.

REFERENCES

Manzoor, A., et al. (2024). Customer Churn Prediction: A Systematic Review of Recent Advances in Machine Learning. Applied Sciences, 14(6), 2501.

Jain, H., Yadav, G., & Manoov, R. (2024). Bank Customer Churn Prediction Using SMOTE: A Comparative Analysis. Qeios Journal of Engineering, 12(4).

Ehsani, F., & Hosseini, M. (2025). Customer Churn Prediction in Digital Banking: A Comparative Study of XAI Techniques. International Journal of Information Management Data Insights, 5(1), 100-112.

Plotnikova, V., Dumas, M., & Milani, F. (2022). Designing a data mining process for the financial services domain. Journal of Financial Data Science, 4(3), 1-22.