Artificial Intelligence - Supervised Learning

Second Project

Banking Customer Churn Prediction

Group A2_33

Alberto Serra   João Alves   Eduardo Sousa

Problem Specification Review

Predicting Customer Churn with Supervised Learning

The Issue and main goal

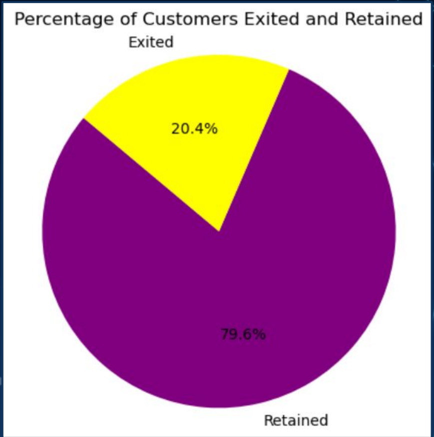
Issue - Customer churn impacts revenue and growth in banking.
Goal - Identify high-risk customers using predictive modeling.

Key Attributes

Demographics - Age, Geography, Gender
Financials - Balance, CreditScore, NumOfProducts
Engagement - HasCrCard, IsActiveMember, Tenure
Outcome - Churn status (Exited)

The Dataset

10,000 customers
14 attributes
Churn outcome
20.4% churn rate



Model Training

Applied various **classification algorithms** (e.g., SVM, Logistic Regression, Random Forest)
Evaluated models using **cross-validation** (F1 score, accuracy)
Selected **CatBoost** as the best model based on performance

Pre-processing and cleaning

Removed identifiers
Encoded categories
Normalized data
Filtered out salary values
Checked for missing values
Addressed outliers

Related Work and References

01

Course Slides

Theoretical **class slides** about machine learning topics

02

Online Documentation and Tutorials

Python Machine Learning libraries documentation (Scikit-Learn, TensorFlow, PyTorch). Programming tutorials and code snippets from **GeeksforGeeks** and **Stack Overflow**.

03

AI Tools in Practice

Coding strategies improved by AI technologies like **ChatGPT**, **Gemini** and **GitHub Copilot**;

04

Research and Case Studies

Analysis of academic papers on churn prediction models from platforms like **Google Scholar** and **ResearchGate**.

05

Open Source Contributions

Machine learning communities on **Reddit** and **GitHub** for code examples, discussions and feedback.

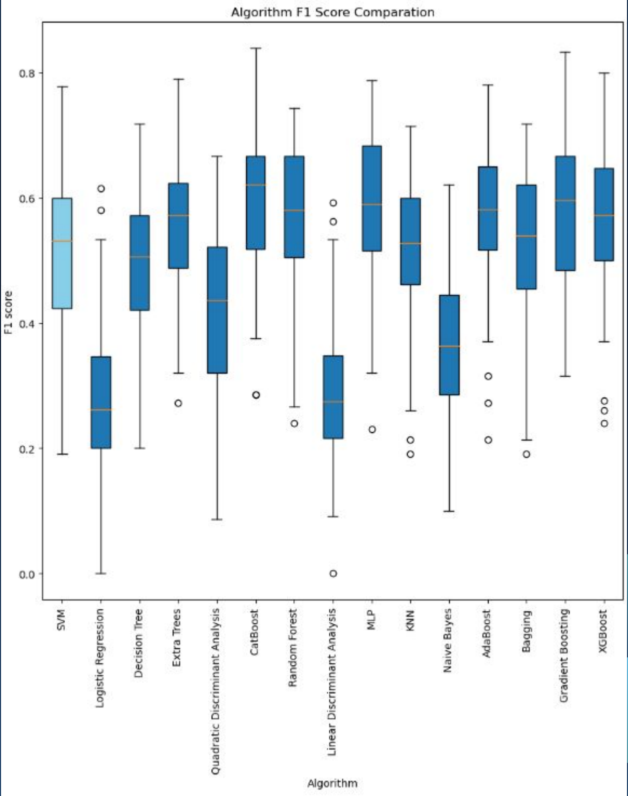
Tools and Algorithms for Churn Prediction

- ***Data Preparation and Visualization***
  - Pandas, NumPy
  - Matplotlib, Seaborn

- ***Machine Learning Models***
  - **Baseline:** Logistic Regression
  - **Tree-Based:** Decision Tree, Random Forest
  - **Ensemble:** Gradient Boosting, AdaBoost, Extra Trees, CatBoost, XGBoost
  - **Support Vectors:** SVM
  - **Neural Networks:** MLPClassifier
  - **Probabilistic:** Naive Bayes, Linear and Quadratic Discriminant Analysis

- ***Model Evaluation and Feature Processing***
  - Accuracy, Precision, Recall, F1 Score, ROC Curves
  - Label Encoding, Standard Scaling

- ***Comparative Analysis***
  - Cross-validation, Boxplots for Algorithm Comparison
  - **Boxplots for Algorithm Comparison:** Visualized F1 Score and Accuracy for different models.



Data Preprocessing and EDA

Data Cleaning

Dropped irrelevant columns	Removed **RowNumber, CustomerId, Surname**
Filtered salary data	Removed rows with **EstimatedSalary** below 1000
Identified outliers	Detected and removed **EstimatedSalary** outliers between 1000 and 5000
Saved cleaned data	Exported to **cleaned_churn_data.csv**

Exploratory Data Analysis (EDA)

Initial exploration	Displayed first few rows and **basic statistics**
Missing values check	Ensured **no missing values**
Class distribution visualization	Created pie chart for Exited (**churn vs. non-churn**)
Pairplot analysis	Visualized **relationships** between features.
Categorical feature analysis	Used **count plots** for **Geography, Gender, HasCrCard, IsActiveMember, NumOfProducts**
Numerical feature analysis	Utilized box plots for **CreditScore, Age, Balance, EstimatedSalary, Tenure**

Data Analysis – Correlation and Insights

Correlation Analysis

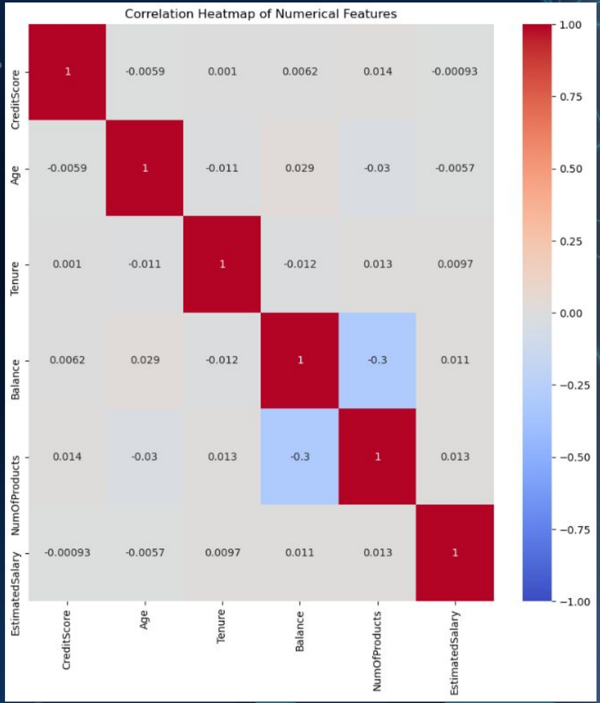
- *Heatmap*
  - Visualized **correlation matrix** to identify significant correlations
- *Key Correlations*
  - **Age and Churn:** Older customers more likely to churn.
  - **Balance and Churn:** Higher balances show varied churn rates.
  - **Number of Products and Churn:** Fewer products correlate with higher churn.

Data Splitting

- *Data split into training and testing sets*
  - **Training set:** 80% of data.
  - **Testing set:** 20% of data.
  - Ensures the model is trained on one part and evaluated on another.

Insights from Data Analysis

- **Credit Score:** Not a strong predictor of churn; lower median for exited customers.
- **Age:** Older customers are more likely to churn; higher median age for exited customers.
- **Balance:** Higher balances correlate with increased churn; higher median balance for exited customers.
- **Number of Products:** Fewer products lead to higher churn.
- **Tenure:** No clear pattern with churn; no significant difference between exited and retained customers.
- **Estimated Salary:** No significant correlation between exited and retained customers.



Model Selection and Evaluation

Model Selection

1

Classification Algorithms Considered:
Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, Extra Trees, Quadratic Discriminant Analysis, CatBoost, Linear Discriminant Analysis, MLP, KNN, Naive Bayes, AdaBoost, Bagging, Gradient Boosting, XGBoost

Cross-Validation

2

K-Fold Cross-Validation:
Used 90 splits for robustness.

Evaluation Metrics:
F1 Score
Accuracy

Data Preparation

3

Feature Types:
Categorical and Continuous - CreditScore, Age, Tenure, Balance, NumOfProducts, EstimatedSalary, Geography, Gender, IsActiveMember, HasCrCard.

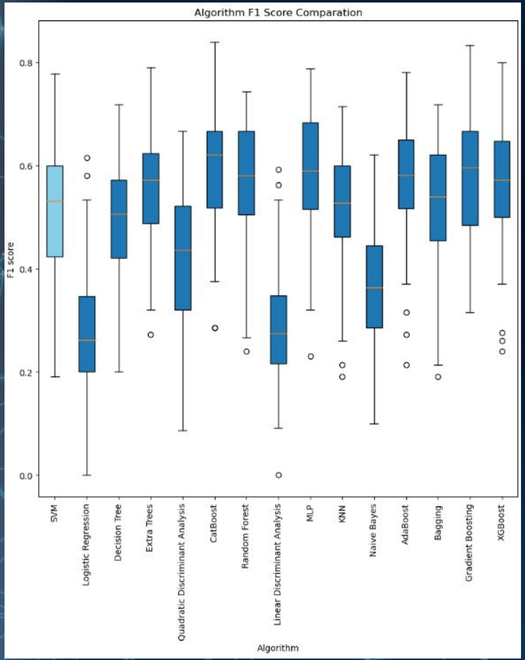
Data Splitting:
Training (80%) and Testing (20%) subsets.

Encoding:
Used LabelEncoder for categorical variables.

Normalization:
Applied StandardScaler to numerical features.

Model Performance – F1 Score, Accuracy

Evaluation Metrics

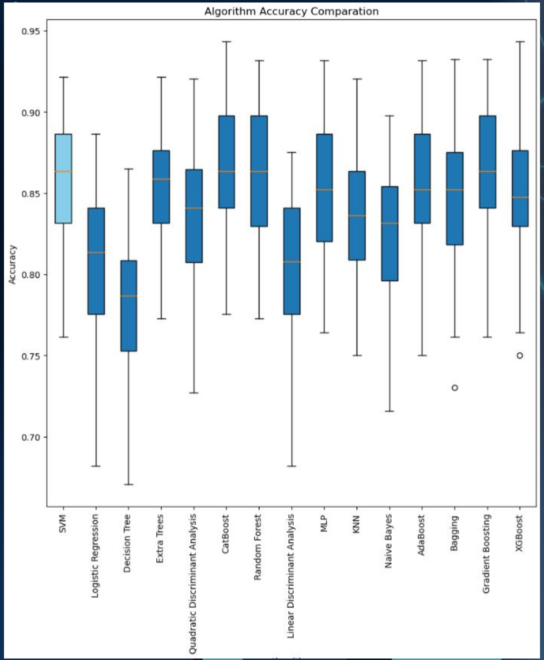


F1 Score

- Measures both **precision and recall.**
- Useful for evaluating **imbalanced datasets.**
- **Best performing models** with highest median **F1 scores** - CatBoost, Random Forest, Gradient Boosting
- **CatBoost** - best performance with a high median F1 score, indicating its robustness.

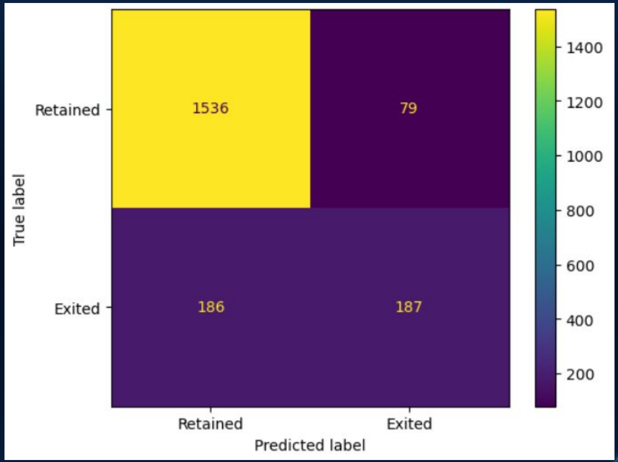
Accuracy

- **Ratio of correctly predicted** instances to total instances
- Effective metric for **balanced datasets**
- **Best performing models** with highest median accuracy scores - SVM, Random Forest, Gradient Boosting, CatBoost
- Consistently **high accuracy** - SVM, CatBoost



Best Model and Results

Best Model Selection	CatBoost based on **overall performance metrics** (F1 score and accuracy) and execution time.
Training and Testing Performance	**Training accuracy:** 90.47%, **Testing accuracy:** 86.67%
Confusion matrix for CatBoost	**Precision, recall, and F1 scores** for both retained and exited classe
Precision, Recall, F1 Score	**Retained:** Precision = 0.89, Recall = 0.95, F1 Score = 0.92 **Exited:** Precision = 0.70, Recall = 0.50, F1 Score = 0.59



Results Conclusion

- CatBoost showed the **highest performance**.
- Identified retained customers with **high precision and recall**.
- Despite **lower accuracy** for exited customers, its **robustness** makes it the best choice.

Conclusion and Insights

Key Predictors, Model Performance, and Practical Applications in Banking

- ***Significant Predictors:***
  - **Higher churn** in Germany.
  - **Higher churn** among females.
  - **Lower churn** among active members.
  - **Higher churn** with fewer products.
- ***Top Performing Models***
  - **CatBoost:** Best overall with high precision and recall.
  - **Random Forest & Gradient Boosting:** Consistent high accuracy and robustness.
- ***Worst Performing Models:***
  - **Naive Bayes:** Lower accuracy and higher error rates.
  - **KNN:** Poor performance in handling high-dimensional data.
  - **Decision Tree:** High variance and overfitting issues.
- ***Top Metrics***
  - **Training Accuracy:** 90.47%
  - **Testing Accuracy:** 86.67%
  - **High precision and recall** for retained customers.
  - **F1 Score:** 0.92 (retained customers), 0.59 (exited customers)
  - **Precision:** 0.89 and Recall: 0.95 (for retained customers)
- ***Real-Life Applicability:***
  - Integrate into banking CRM systems to **predict and reduce churn**.
  - Implement personalized **retention strategies** (e.g financial advice, targeted promotions)
  - Enhances **customer satisfaction and loyalty**.
  - Demonstrates **potential of predictive analytics** in the banking industry.