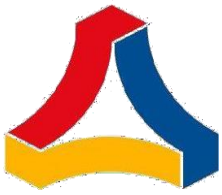

TOBB ETÜ

Customer Churn Prediction Using Machine Learning Models in the Banking Sector

YAP 470

Aytuğ Şahinkanat

Bektaş Batuhan Kestek

Remzi Mert Tehneldere

1. Abstract

Customer churn in the banking sector is a critical issue with a direct impact on profitability. This study proposes a data-driven approach to predict customer churn. By combining demographic data, customer behaviors, and transaction histories, this work specifically focuses on **time-windowed feature engineering**. These features, derived to capture recent changes in customer behavior, are expected to enhance predictive power. The study will compare the performance of advanced gradient boosting models, such as **XGBoost** and **CatBoost**—which are optimized for class imbalance and capable of modeling non-linear relationships—against a baseline Logistic Regression model

2. Related Work

Previous research on churn prediction in the banking sector has highlighted the importance of transaction-based feature engineering, similar to **RFM (Recency, Frequency, Monetary)** analyses. Recent studies, in particular, emphasize that behavioral changes (e.g., a sudden drop in recent transaction frequency) serve as early indicators of potential churn.

In the literature, Logistic Regression, Random Forest, and gradient boosting algorithms such as XGBoost are commonly used to address this problem. This work aims to extend the existing literature by incorporating time-windowed average transaction features, specifically covering the last 1, 3, and 6-month periods, and to examine their impact on model performance.

3. Methodology

This section details the data collection process, the applied feature engineering techniques, and the planned modeling approach.

3.1. Data Collection and Merging

Three main datasets were used in this study. These sets were merged based on the customer ID (cust_id):

1. **Demographic Data** (customers.csv)
2. **Churn Labels and Reference Dates** (referance_data.csv)
3. **Full Transaction History** (customer_history.csv)

The merging process resulted in a clean primary dataset containing 133,287 unique customers and 30 features.

3.2. Exploratory Data Analysis (EDA)

The initial exploratory data analysis (EDA) revealed that the distribution of customer churn **does not have a linear relationship** with key demographic and behavioral attributes, such as age, employment type, and current product usage. This finding supports the necessity of using non-linear models.

3.3. Feature Engineering

Two primary feature engineering approaches were implemented to enhance model performance:

1. **Time-Windowed Features:** To capture current changes in customer behavior, metrics such as average transaction amount and transaction frequency were calculated for the **last 1, 3, and 6-month** time windows for each customer, looking back from their reference date. Missing values were imputed using each customer's own overall average.
2. **RFM Analysis (Recency, Frequency, Monetary):** RFM analysis, frequently used in the literature, was also tested as a feature set in this study.
    - **Recency:** When the customer made their last transaction.
    - **Frequency:** How often the customer transacted in a given period.
    - **Monetary:** The total monetary value of the customer's transactions.

3.4. Feature Selection and Correlation Analysis

In the feature selection process, the focus was on removing variables with high correlation (correlation > 0.67) to prevent issues of **multicollinearity**.

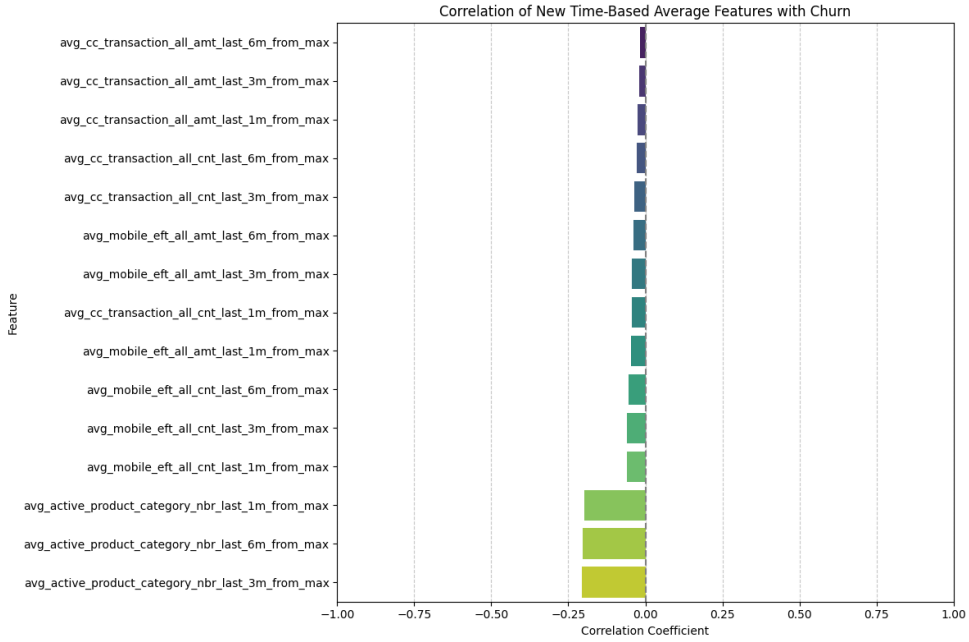

Image 1 Correlation Graph of numerical features

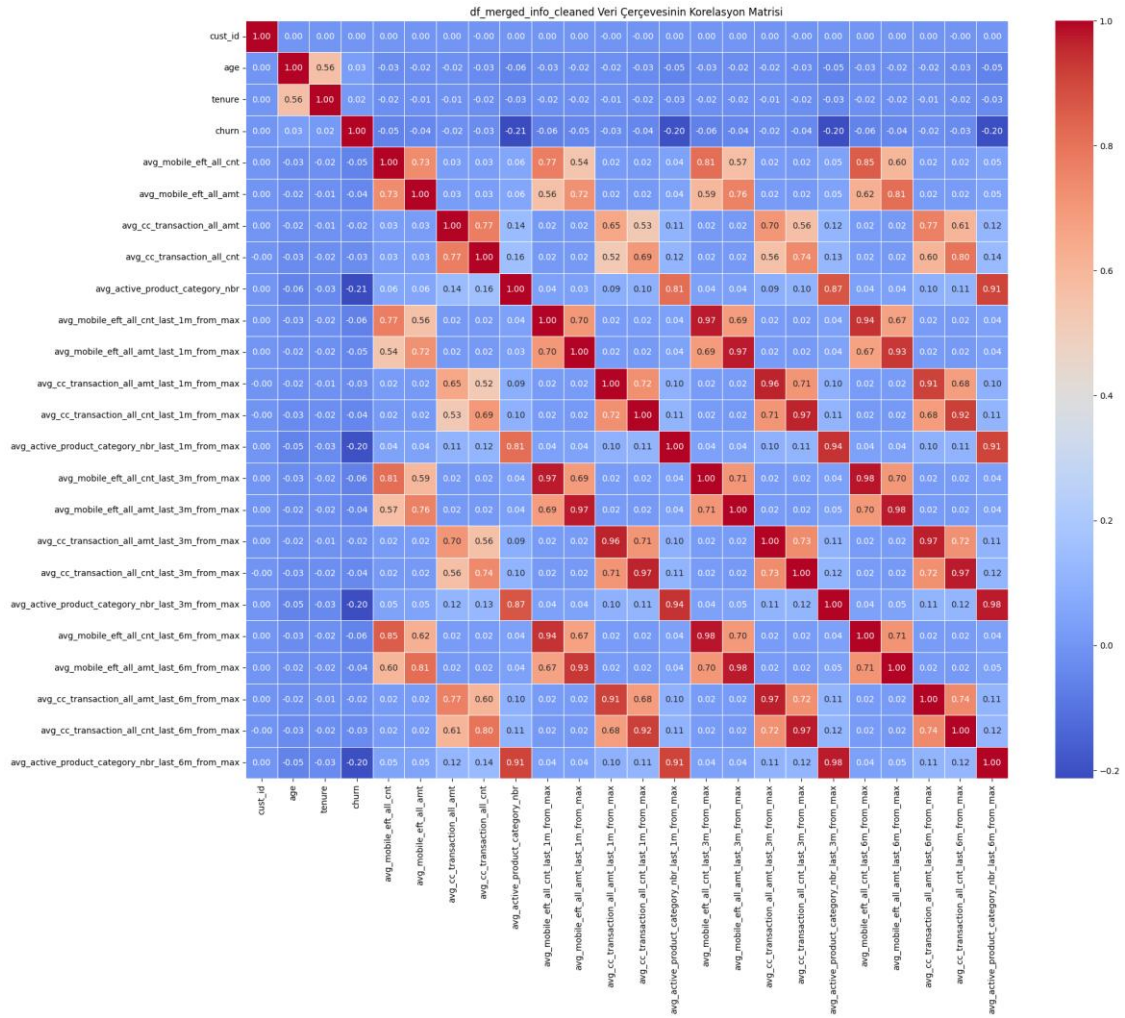

Image 2 Correlation Heatmap

During the feature selection process, the objective was to identify a subset of features that demonstrated the highest correlation with the target variable (churn) while simultaneously mitigating multicollinearity. To achieve this, a strict threshold was applied, and any variables exhibiting an inter-correlation higher than 0.67 were systematically removed. The final, optimized set of features retained for modeling consists of the following: avg_active_product_category_nbr, avg_mobile_eft_all_cnt_last_1m_from_max, avg_cc_transaction_all_cnt_last_1m_from_max, avg_mobile_eft_all_amt_last_6m_from_max, age, tenure, and avg_cc_transaction_all_amt.

Furthermore, the correlations between the **RFM metrics** (tested during feature engineering) and the target variable (churn) were examined.

```
rfm_churn_cols = ['Recency', 'Frequency', 'Monetary', 'churn']
correlation_matrix = df_rfm_combined[rfm_churn_cols].corr(method='pearson')

churn_correlations = correlation_matrix['churn'][:-1]

print("Correlation of RFM features with Churn:")
print(churn_correlations)
```

```
Correlation of RFM features with Churn:
Recency      0.025325
Frequency   -0.053285
Monetary    -0.053285
Name: churn, dtype: float64
```

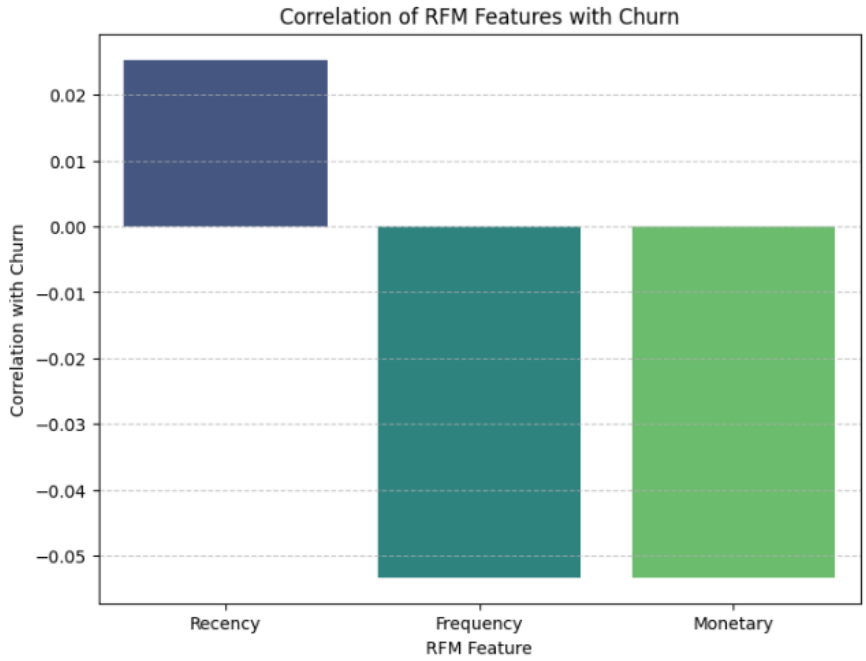

Image 3,4 Correlation Graph of RFM features with Churn

The analysis revealed that the **RFM features had statistically very low correlations (e.g., < 0.055) with the churn variable.** Due to this weak relationship, it was determined that the RFM features would not provide a significant contribution to the model's predictive power and could potentially introduce noise; therefore, these features were excluded from the final model set.

Furthermore, this analysis was extended to key categorical (non-numeric) features to understand their impact on churn status. The distribution of churn rates was analyzed across different groups, such as employment type, customer tenure brackets, and the number of active products. These findings confirmed that churn behavior is non-linearly distributed across these categorical attributes as well.

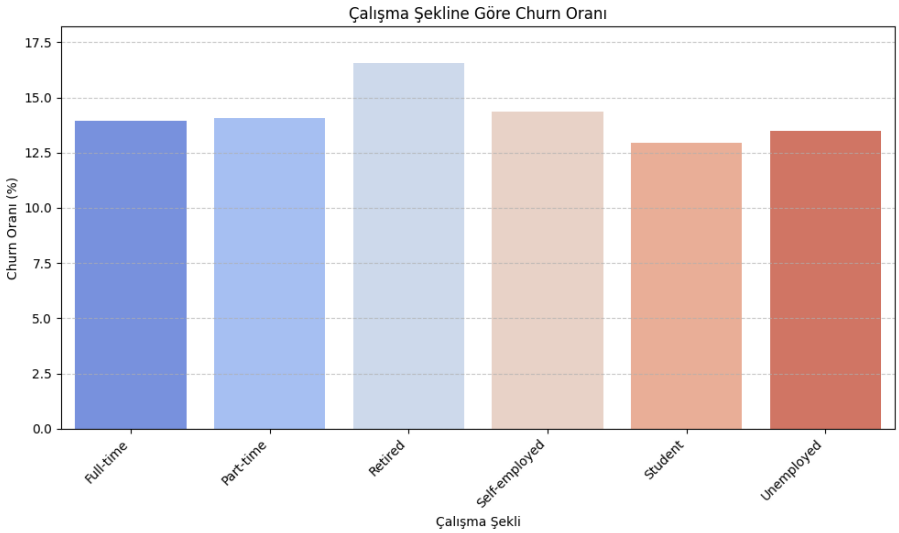

Image 5 Graph of distribution of working sector with churn

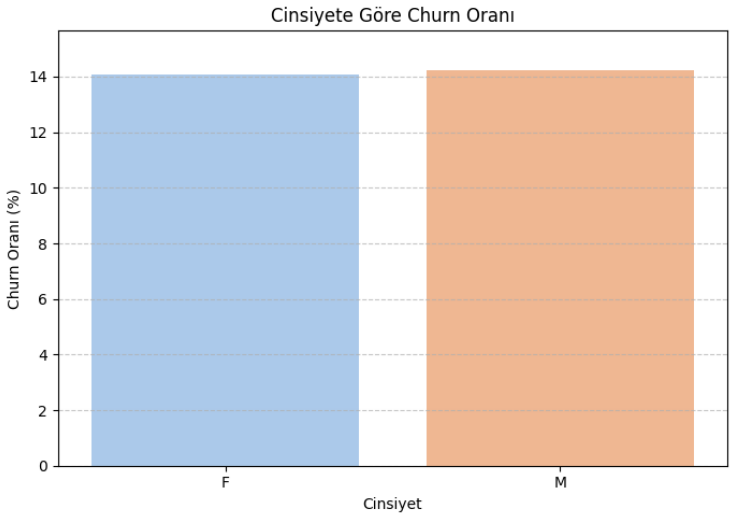

Image 6 Graph of distribution of sex with churn

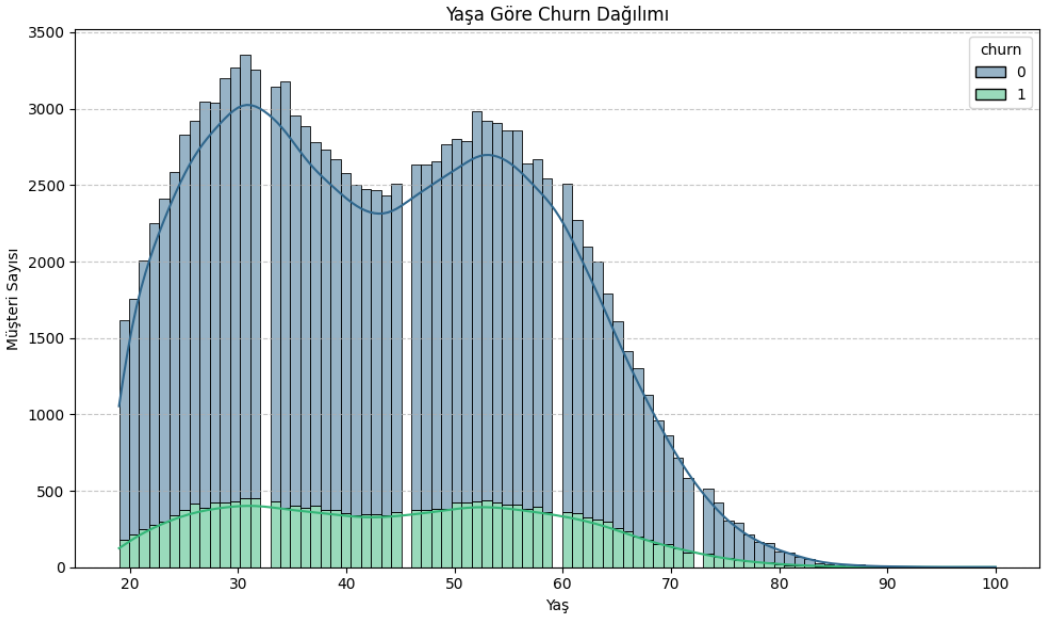

Image 7 Graph of distribution of age with churn

As can be concluded from these graphs, the non-numeric (categorical) data did not demonstrate a significant impact on customer churn.

4. Modeling and Evaluation Plan

4.1. Planned Models

To effectively model the non-linear structure observed in the EDA, the use of the following algorithms is planned:

1. **Logistic Regression (Baseline):** To be used to establish baseline performance as a linear model.
2. **XGBoost (Extreme Gradient Boosting):** A high-performance, scalable gradient boosting model.
3. **CatBoost:** A modern gradient boosting model particularly successful at handling categorical variables, often requiring less hyperparameter tuning.

4.2. Model Validation and Performance Metrics

The dataset will be split into **80% training** and **20% testing** sets to train and evaluate the models.

5-fold cross-validation will be used on the training data for hyperparameter tuning and to measure the model's generalization capability.

Given the **class imbalance** problem common in churn data, model performance will not only be assessed by accuracy but will be comprehensively evaluated using the following metrics:

- **Precision**
- **Recall**
- **F1-Score**
- **ROC-AUC (Area Under the Receiver Operating Characteristic Curve)**

4.3. Future Work (Interpretability)

After training the models, the goal is not only to achieve high accuracy but also to ensure they do not remain "black boxes." To this end, a model interpretability analysis using **SHAP (SHapley Additive exPlanations)** values is planned to understand which features are the key drivers of churn and to provide actionable insights to business units.