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 (Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2018). RFM ranking – An effective approach to customer segmentation. Decision Analytics, 5(1), 1-20.)

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.

After the research, the models like XGBoost and CatBoost are determined to use for the nonlinear data.

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):

- customers.csv (cust_id, gender, age, province, religion, work_type, work_sector, tenure)
- reference_data.csv (cust_id, ref_date, churn)
- customer_history.csv (cust_id, date, mobile_eft_all_cnt, active_product_category_nbr, mobile_eft_all_amt, cc_transaction_all_amt, cc_transcation_all_cnt)

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.

Numerical Features and Churn Correlations(from Customer.csv & reference.csv)

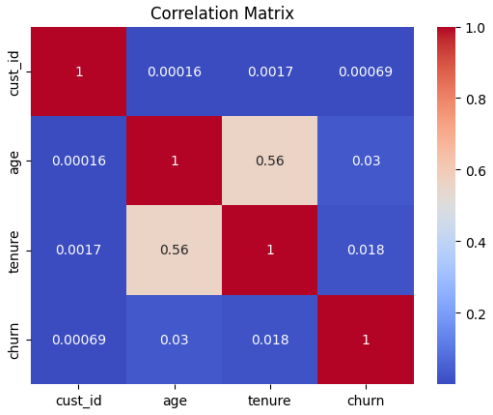


Image3.2.1: Correlations of Numerical Features

Categorical Features Churn Rates(from Customer.csv & reference.csv)

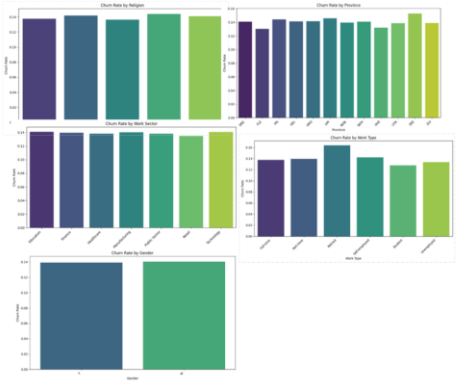


Image3.2.2: Churn Ratings of the categorical data

3.3. Feature Engineering

As a beginning of the feature engineering, the average features have been considered from Customer History data. And then the Reference data, Customer Information data and the Customer History data (mean values of the history data) have been merged into a datafram

```
   #   Column                        Non-Null Count   Dtype
  ---  ------                        --------------   -----
   0   cust_id                       133287 non-null  int64
   1   gender                        133287 non-null  object
   2   age                           133287 non-null  int64
   3   province                      133287 non-null  object
   4   religion                      133287 non-null  object
   5   work_type                     133287 non-null  object
   6   work_sector                   110520 non-null  object
   7   tenure                        133287 non-null  int64
   8   ref_date                      133287 non-null  datetime64[ns]
   9   churn                         133287 non-null  int64
   10  avg_mobile_eft_all_cnt        130622 non-null  float64
   11  avg_mobile_eft_all_amt        130622 non-null  float64
   12  avg_cc_transaction_all_amt    129304 non-null  float64
   13  avg_cc_transaction_all_cnt    129304 non-null  float64
   14  avg_active_product_category_nbr  133287 non-null  float64
  dtypes: datetime64[ns](1), float64(5), int64(4), object(5)
```

e.

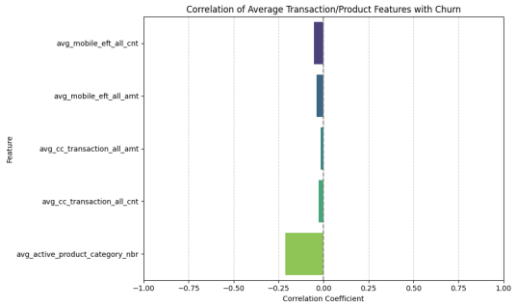
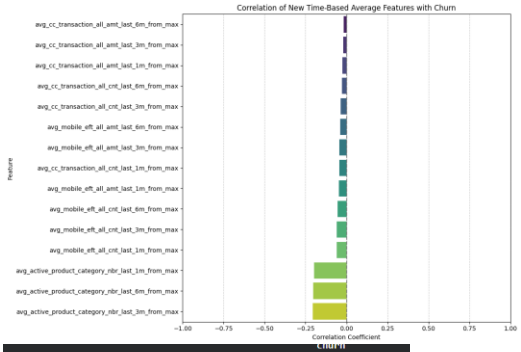


Image 3.3.1: Customer Information & Reference data & Customer History(mean) .

Image 3.3.2: The correlations between mean features and Churn.

Because of the last months are essentially important with determining churn, the time windowed features have been considered. 1, 3 and 6 month average features have been considered.



avg_mobile_eft_all_cnt_last_1m_from_max	-0.060908
avg_mobile_eft_all_amt_last_1m_from_max	-0.046541
avg_cc_transaction_all_amt_last_1m_from_max	-0.025502
avg_cc_transaction_all_cnt_last_1m_from_max	-0.043600
avg_active_product_category_nbr_last_1m_from_max	-0.198665
avg_mobile_eft_all_cnt_last_3m_from_max	-0.058627
avg_mobile_eft_all_amt_last_3m_from_max	-0.043404
avg_cc_transaction_all_amt_last_3m_from_max	-0.021161
avg_cc_transaction_all_cnt_last_3m_from_max	-0.035008
avg_active_product_category_nbr_last_3m_from_max	-0.204989
avg_mobile_eft_all_cnt_last_6m_from_max	-0.055858
avg_mobile_eft_all_amt_last_6m_from_max	-0.039737
avg_cc_transaction_all_amt_last_6m_from_max	-0.017787
avg_cc_transaction_all_cnt_last_6m_from_max	-0.028831
avg_active_product_category_nbr_last_6m_from_max	-0.203048

Image 3.3.3&4: Correlations between churn and new time windowed features

Plenty number of features have been tested. 3 of these are about inactivity. 'total_inactive_months', 'max_consecutive_inactive_months', 'inactive_days_last_6m' are three of these features. Due to the low correlations, these have been eliminated.

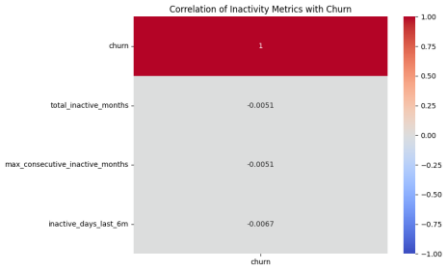


Image 3.3.3&4: Correlations between churn and new time windowed features

Regarding to the RFM analysis , the frequency and monetary features have been constructed. Due to the churn referencing in the competition, Recency has more important than the Monetary and Frequency features. Even though the Frequency and Monetary should be constructed and be analyzed due to their churn status.

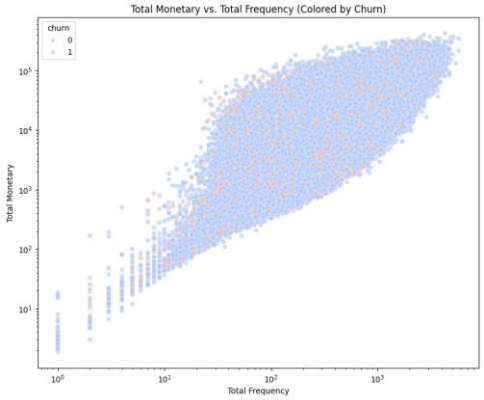


Image 3.3.5: Frequency and Monetary, colored by Churn

As it seen, the Monetary and Frequency data is irrelevant for churn status. These data do not provide any effective information about churn status.

Other visuals for determining any binary relationship between these numerical features.

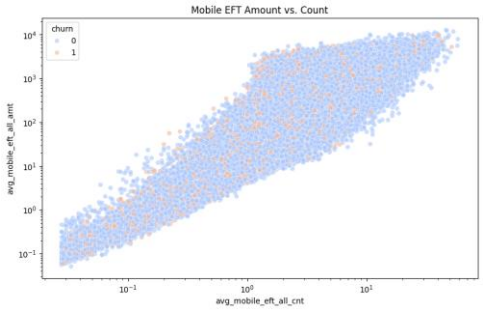


Image 3.3.6: Mobile EFT Amount and Mobile EFT Count, colored by Churn

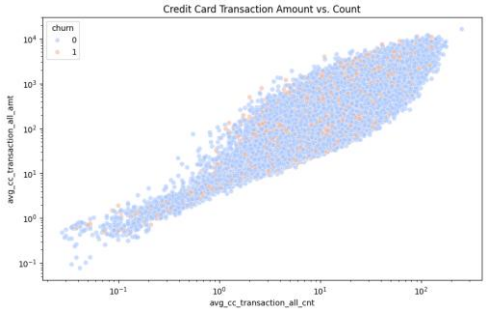


Image 3.3.7: Mobile EFT Amount and Mobile EFT Count, colored by Churn

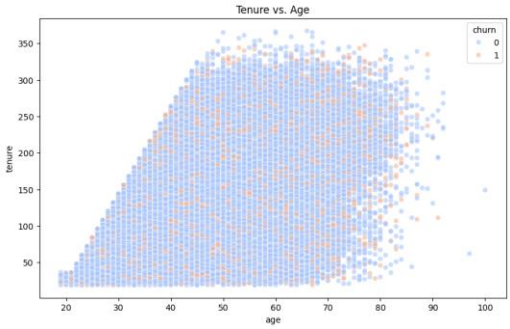


Image 3.3.8: Mobile EFT Amount and Mobile EFT Count, colored by Churn

After detailed analysis about these features, it is considered that there are not significant relationships between features and considered using nonlinear models like XGBoost because of the weak relations.

XGBoost has powerful key features for nonlinear churn analysis. Its like L1 and L2 regularization for preventing the overfiting. Gradient descent for minimize loss during the iterations. And it handles missing values.

3.4. Feature Selection and Reduction

3.4.1. Numerical Feature

After having plenty of uncorrelated new features, they have been reduced.

'total_inactive_months', 'max_consecutive_inactive_months', 'inactive_days_last_6m' have been reduced because of the low correlation with churn(<0.01, due to Image 3.3.3&4)

Total Frequency and Total Monetary features are unsuccessful due to the colored churn graph (Image 3.3.5). It is not determined any relations between these features with churn status.

After the elimination due to the correlations, the PCA cumulative explained variance graph has been constructed with PCA from sklearn.decomposition. It is found that the number of features that needed to explain the variance is 9. It simplifies the model, removes **multicollinearity** (the 80% correlation issue you mentioned), and helps the XGBoost model train

faster by using fewer, more meaningful variables.

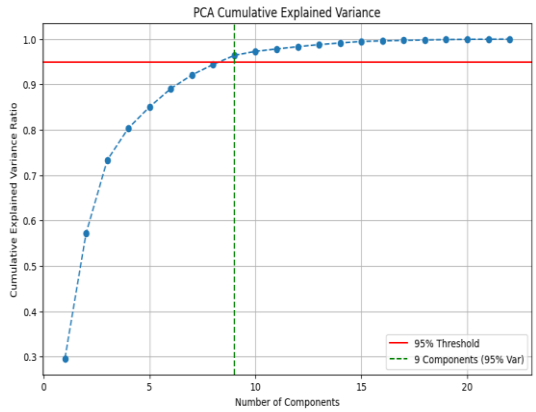


Image 3.4.1: PCA cumulative explained variance

To streamline the model and improve performance, a feature selection strategy was implemented based on multicollinearity and target relevance. Specifically, for any pair of features with a correlation higher than 80%, the feature that has a lower correlation with the target variable (Churn) is removed.

```
*** Features identified for dropping due to high inter-correlation and lower correlation with churn:
    - avg_cc_transaction_all_cnt_last_3m_from_max
    - avg_mobile_eft_all_cnt_last_3m_from_max
    - avg_mobile_eft_all_cnt_last_6m_from_max
    - avg_mobile_eft_all_amt_last_3m_from_max
    - avg_mobile_eft_all_cnt
    - avg_mobile_eft_all_amt
    - avg_active_product_category_nbr_last_6m_from_max
    - avg_mobile_eft_all_amt_last_6m_from_max
    - avg_cc_transaction_all_amt_last_3m_from_max
    - avg_cc_transaction_all_cnt_last_6m_from_max
    - avg_active_product_category_nbr_last_3m_from_max
```

Image 3.4.2: Eleminated Features

- Multicollinearity Removal: Eliminating features that repeat the same information.
- Target-Driven Selection: Prioritizing variables that have a stronger relationship with Churn.
- Dimensionality Reduction: Simplifying the input space to improve the performance of the XGBoost model.

3.4.1. Categorical Features

Churn rates of categorical features have been analyzed. An analysis of the churn rates for categorical features was conducted. Since the **gender** and **religion** features were found to have almost no impact on churn, they were excluded from the analysis.

Global Churn Rate: 0.1416

Churn Rates by Gender M: 0.1422, **F:** 0.1409

Churn Rates by Province ZEE: 0.1530, **LIM:** 0.1478, **FRI:** 0.1451, **GEL:** 0.1431, **NOH:** 0.1427, **DRE:** 0.1427, **GRO:** 0.1418, **NOB:** 0.1411, **ZUI:** 0.1404, **UTR:** 0.1403, **FLE:** 0.1334, **OVE:** 0.1319

Churn Rates by Religion O: 0.1462, **J:** 0.1433, **U:** 0.1425, **C:** 0.1392, **M:** 0.1380

Churn Rates by Work Type Retired: 0.1655, **Self-employed:** 0.1435, **Part-time:** 0.1408, **Full-time:** 0.1395, **Unemployed:** 0.1346, **Student:** 0.1293

Churn Rates by Work Sector Manufacturing: 0.1432, **Education:** 0.1422, **Technology:** 0.1412, **Public Sector:** 0.1405, **Finance:** 0.1401, **Healthcare:** 0.1400, **Retail:** 0.1357

3.5 Model Evaluation

Based on our exploratory data analysis, the dataset exhibits a clear non-linear structure, making traditional linear models unsuitable for capturing the underlying patterns. After implementing a feature selection strategy to remove redundant variables with high inter-correlation (over 80%) and low target relevance, we identified the key features shown below:

```
Updated df_merged_info_cleaned info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 133287 entries, 0 to 133286
Data columns (total 13 columns):
 #   Column                                  Non-Null Count   Dtype
---  ------                                  --------------   -----
 0   cust_id                                 133287 non-null  int64
 1   age                                     133287 non-null  int64
 2   province                                133287 non-null  object
 3   work_type                               133287 non-null  object
 4   work_sector                             110520 non-null  object
 5   ref_date                                133287 non-null  datetime64[ns]
 6   churn                                   133287 non-null  int64
 7   avg_cc_transaction_all_cnt              129304 non-null  float64
 8   avg_active_product_category_nbr         133287 non-null  float64
 9   avg_mobile_eft_all_cnt_last_1m_from_max 133287 non-null  float64
 10  avg_mobile_eft_all_amt_last_1m_from_max 133287 non-null  float64
 11  avg_cc_transaction_all_amt_last_1m_from_max  133287 non-null  float64
 12  avg_cc_transaction_all_cnt_last_1m_from_max  133287 non-null  float64
dtypes: datetime64[ns](1), float64(6), int64(3), object(3)
memory usage: 13.2+ MB
```

Image 3.5.1: Remaining Features

Model Selection Strategy

Given the non-linear nature of the data and the complexity of the feature interactions, we decided to employ advanced Gradient Boosting architectures rather than simpler ensemble methods like Random Forest. Our training approach focuses on two primary models:

XGBoost (eXtreme Gradient Boosting): Selected for its high efficiency and ability to handle non-linear relationships through optimized gradient descent and regularization.

```
xgb_model = XGBClassifier(
    objective='binary:logistic',  # Bina
    eval_metric='auc',            # Eval
    use_label_encoder=False,      # Supp
    random_state=42,              # For
    n_estimators=500,             # Numb
    learning_rate=0.1,            # Step
    max_depth=6,                  # Maxi
    subsample=0.8,                # Subs
    colsample_bytree=0.8,         # Subs
    scale_pos_weight=scale_pos_weight, #
    tree_method='hist'            # Use
)
```

Image 3.5.2: XGBoost Model Parameters

CatBoost: Utilized specifically for its superior handling of categorical variables and its robust performance against overfitting in complex datasets.

```
cat_model_weighted = CatBoostClassifier(
    iterations=500,  # Number of boosting rounds
    learning_rate=0.1, # Step size shrinkage to prevent overfitting
    depth=6,         # Depth of the trees
    loss_function='Logloss', # For binary classification
    eval_metric='AUC', # Metric to monitor during training
    random_seed=42,
    verbose=100,     # Print progress every 100 iterations
    early_stopping_rounds=50, # Stop if AUC on validation set doesn't improve for 50 rounds
    class_weights=class_weights_dict # Apply class weights
)
```

Image 3.5.3: CatBoost Model Parameters

Rationale: While Random Forest builds trees independently through bagging, XGBoost and CatBoost use boosting to learn sequentially from previous errors. This iterative optimization is significantly more effective at minimizing the loss function in non-linear scenarios, ultimately leading to higher predictive accuracy for Churn.

3.5.1 XGBoost Results

XGBoost Model Evaluation on Test Set:

- Accuracy: 0.6627
- Precision: 0.2294
- Recall: 0.5859
- F1-Score: 0.3297
- ROC AUC: 0.6898

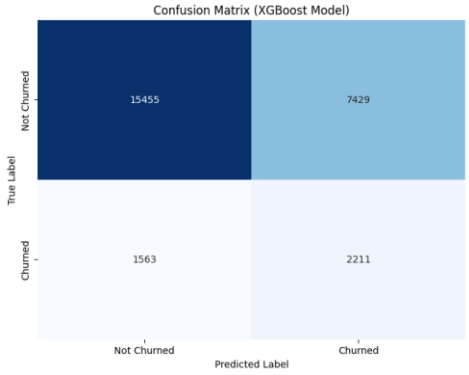


Image 3.5.1.1: Confusion Matrix

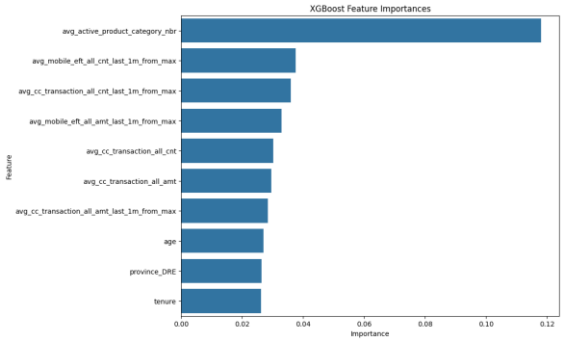


Image 3.5.1.2: XGBoost Feature Importances

3.5.2 CatBoost Results

CatBoost Model Evaluation on Test Set:

- Accuracy: 0.5881
- Precision: 0.2216
- Recall: 0.7602
- F1-Score: 0.3432
- ROC AUC: 0.7134

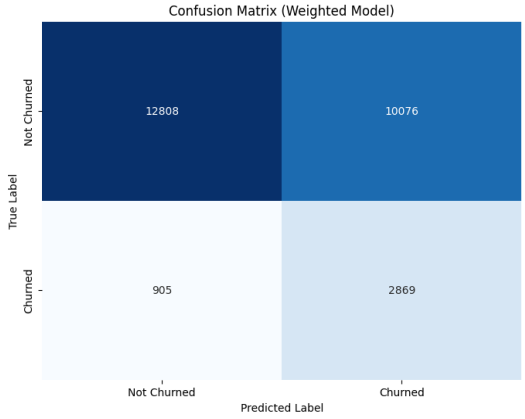


Image 3.5.2.1: Confusion Matrix

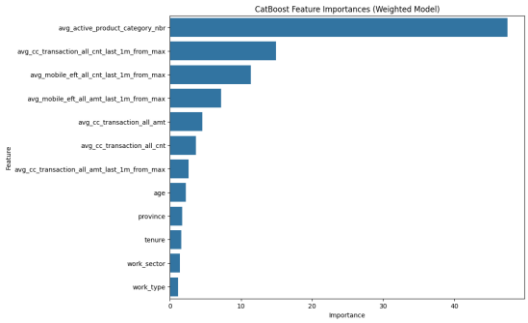


Image 3.5.2.2: Confusion Matrix

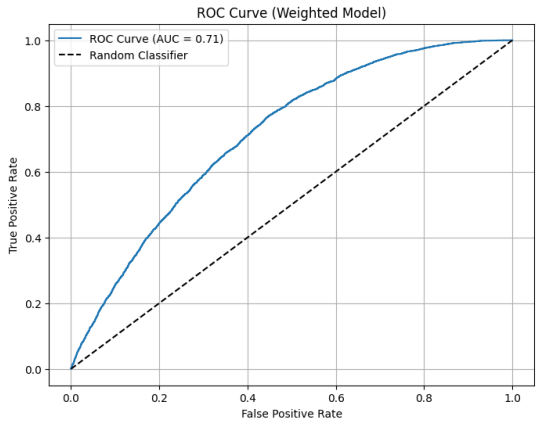


Image 3.5.2.3: ROC Curve for CatBoost Model

4 Model Comparison and Performance Evaluation

The performance of the predictive models was evaluated using confusion matrices and key classification metrics, comparing the XGBoost and Weighted CatBoost architectures. While the XGBoost model achieved a higher overall Accuracy (0.6627), it fell short in identifying actual churners, correctly capturing only 2,211 instances with a Recall of 0.5859. In contrast, the Weighted CatBoost model demonstrated superior diagnostic power for the minority class, correctly identifying 2,869 true positive churn cases. Although this led to a slightly lower Accuracy (0.5881) and Precision (0.2216), the Weighted CatBoost model achieved a significantly higher Recall of 0.7602 and a better ROC AUC of 0.7134, making it the more robust choice for detecting at-risk customers.

The Strategic Importance of Recall in Churn Management

In the context of churn analysis, Recall is the most critical metric because the cost of losing a customer far outweighs the cost of a retention campaign. The Weighted CatBoost model's high recall rate means the company can successfully identify approximately 76% of all potential churners. From a business perspective, this allows the organization to accurately target over three-quarters of at-risk customers with specialized loyalty programs or discount campaigns before they terminate their relationship. Even with an F1-Score of 0.3432, the model's ability to minimize missed churners (False Negatives) ensures that the marketing department can intervene effectively, preventing significant revenue loss and maximizing Customer Lifetime Value (CLV) through data-driven retention strategies.

For future analysis, and better performance, effective features should be constructed at the feature engineering part. These newly developed features should be correlated with the churn status.

5 References

1. RFM Analysis (Foundational Concepts**)**

- **Hughes, A. M. (2006).** *Strategic database marketing: The masterplan for starting and managing a profitable, customer-based marketing program.* McGraw-Hill. (Classic text on RFM scoring and customer segmentation).
- **Wei, J. T., Lin, S. Y., & Wu, H. H. (2010).** A review of the application of RFM model. *African Journal of Business Management*, 4(19), 4199-4206.

2. XGBoost & CatBoost (Algorithm Methodology)

- **Chen, T., & Guestrin, C. (2016).** XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
- **Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018).** CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems (NeurIPS)*, 31. (This explains why CatBoost is superior for categorical data).

3. Churn Prediction & Non-Linear Modeling

- **Burez, J., & Van den Poel, D. (2009).** Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3), 4626-4636.
- **Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015).** A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55, 1-9.
- **Ahmed, A. A., & Maheswari, D. (2017).** Churn prediction on huge datasets using XGBoost. *International Journal of Advanced Research in Computer Science*, 8(3).