Data Mining Project
Customer Churn Prediction

Anh-Nhat Nguyen[2034311], Momin Abdullah[1869371], Faisal Habib[1961913], Nyasha Chagonda[1964010], Onur Can Memis[2040827], and Abdelrhman Ahmed[2025322]

Team 7

1 Introduction

The telecommunications industry (Telco) grapples with high customer churn rates, posing significant retention challenges. To tackle this issue, a machine learning classification project aims to predict customer churn by analyzing historical data from Telco. The Kaggle-sourced dataset encompasses customer demographics, subscribed services, tenure, and other pertinent features. The primary goal is to develop a robust predictive model that accurately identifies customers likely to churn, enabling proactive intervention strategies to reduce churn rates and improve customer retention. By leveraging machine learning algorithms to uncover data patterns, this project seeks to generate actionable insights for Telco, facilitating targeted retention campaigns and enhance customer satisfaction

1.1 Dataset

1.2 Data Source

According to the paper [1], customer churn poses a significant challenge for telecommunications companies as it adversely impacts their profitability. This issue is particularly critical given the saturated nature of the global telecommunications market, which makes retaining existing customers increasingly difficult. To address this problem, we are leveraging the Telco Customer Churn dataset, sourced from Kaggle, comprises 7,043 rows, where each row represents a unique customer. Each row contains 21 columns (attributes), providing detailed information about the respective customer's demographic profile (e.g., gender, age range, partner, and dependents), account details (e.g., tenure, contract type, payment method, paperless billing status, monthly charges, and total charges), subscribed services (e.g., phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies), and tenure with the company. The column metadata describes the specific attributes represented in each column, enabling comprehensive analysis of customer characteristics and their relationship with churn behavior within the telecommunications company.

Source of dataset: [5]

1.3 Data Collection

The dataset was obtained from Kaggle, we could download it onto our computer in the form of *csv*. An overview of dataset attribute is presented in Table 1 below.

2 Methodology

2.1 Data Exploration and Cleaning

The data exploration process involved loading the dataset, removing irrelevant attributes, analyzing dimensions, checking for duplicates and missing values, handling data types, visualizing data distributions, and exploring categorical data. These steps provided a comprehensive understanding of the dataset's structure, quality, and characteristics, facilitating further preprocessing and modeling. The details steps are describing as below:

- Loading Initial Data: Loading the dataset and necessary libraries for data manipulation and visualization.
- Removing Irrelevant Attribute: Removing the 'customerID' column as it is not relevant for analysis or modeling.
- Dimension Analysis: Analyzing the dimensions (number of rows and columns) of the dataset after removing the irrelevant attribute.
- Duplicate and Missing Value Check: Checking for and removing duplicate rows, and verifying the absence of missing values in the dataset.
- Dataset Information: Obtaining a summary of the dataset's attributes, data types, and non-null counts.
- Handling Missing Values in 'TotalCharges' Column: Converting the 'TotalCharges' column to numeric data type and filling missing values with the mean value.
- Data Distribution Visualization: Visualizing the distribution of the target variable ('Churn') and numeric attributes using bar plots and histograms.
- Categorical Data Exploration: Exploring the unique values and value counts for each categorical attribute in the dataset.

During the data exploration phase, we conducted a correlation analysis to identify relationships between features and the target variable, Churn. By applying one-hot encoding to categorical variables, we transformed them into binary indicators, facilitating effective analysis in machine learning models.

The correlation analysis involved calculating correlation coefficients, which quantify the strength and direction of the relationship between each feature and Churn. Positive correlations suggest that as one variable increases, the other tends to increase, while negative correlations indicate an inverse relationship.

To visualize the correlations, we employed a bar chart, plotting the correlation coefficients. This graphical representation, as depicted in Figure 1 below, enabled us to readily identify features that exhibited strong correlations with Churn. For instance, if a feature had a high positive correlation, it implied a

strong positive relationship, influencing the prediction of customer churn. Conversely, a feature with a high negative correlation would suggest a strong inverse relationship, impacting churn prediction in the opposite direction.

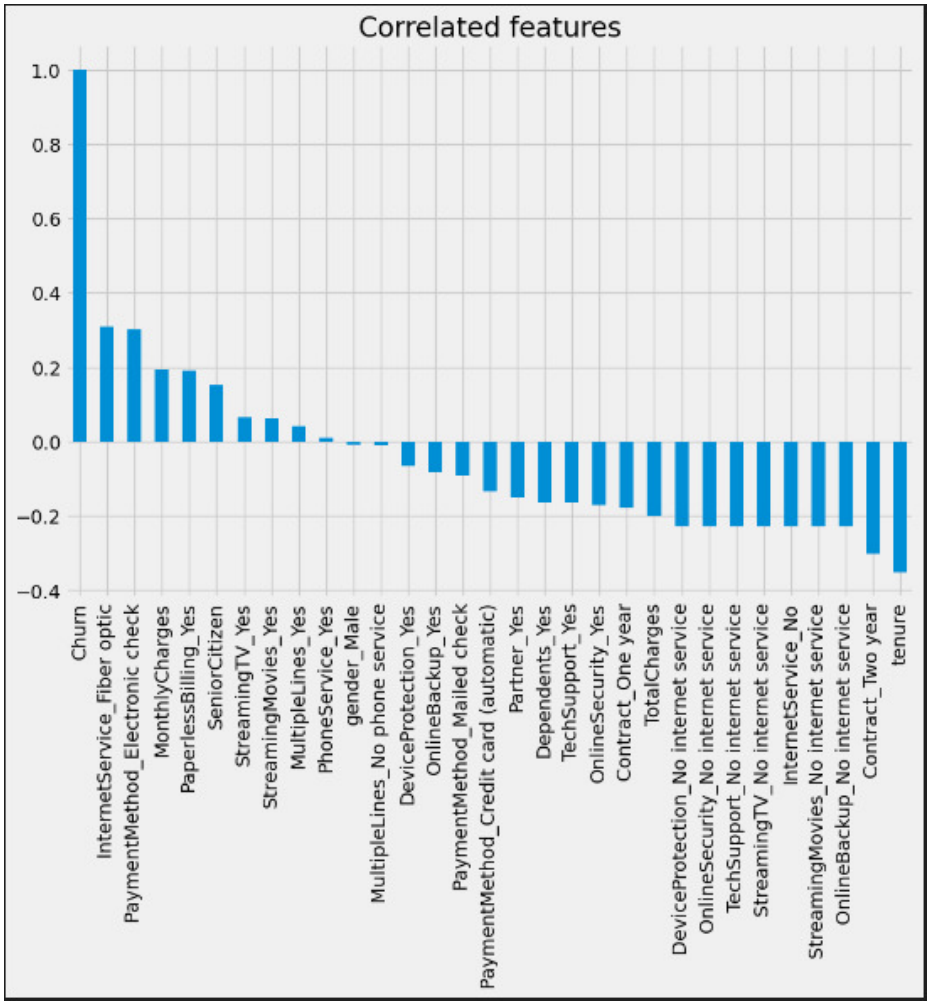

Fig. 1. Features Correlation

By understanding these correlations, we gained valuable insights into the significance of different features in predicting customer churn. This analysis played a crucial role in feature selection and model building, guiding us in constructing more accurate and robust models for churn prediction tasks.

2.2 Data Preprossessing

In the data preprocessing stage, several techniques are employed to prepare the dataset for modeling and analysis. The process involves the following steps:

Data Transformation

- The dataset is divided into numeric and categorical columns for separate processing and modeling approaches.
- The target variable `Churn` is transformed into a binary format suitable for binary classification tasks, facilitating model training and evaluation.

Data and Target Variable Separation

- The dataset is separated into feature variables (`X`) and the target variable (`y`), preparing the data for supervised learning tasks.

Train-Test Split

- The dataset is split into training and testing sets using the `train_test_split` function from scikit-learn.
- Stratified splitting is employed to maintain the class distribution in both the training and testing sets, ensuring representative data for model training and evaluation.

Feature Transformation

- The feature variables (`X`) are transformed using a column transformer created with the `make_column_transformer` function from scikit-learn.
- Categorical columns are encoded using the `OneHotEncoder` to convert them into binary vectors, while numeric columns are scaled using the `MinMaxScaler` to ensure all features are on a similar scale.
- The defined transformer is applied to both the training and testing sets to transform the features accordingly.
- Numeric columns are scaled using Min-Max scaling to normalize their values within the range [0, 1].
- Feature transformation ensures that the data is appropriately processed and standardized, improving model performance and interpretability.

2.3 Experimental setting

This section outlines the methodology employed for predicting customer churn using both machine learning and deep learning models as well as the model selection process. In our machine learning experiments, we evaluated various models to identify the best-performing ones for our text classification task. Among the models tested, an Artificial Neural Network (ANN) served as one of our baseline

models with the highest f1-score. However, to ensure robustness and improve our results, we also explored other algorithms. Among them, the XGBoost classifier emerged as a strong contender. After rigorous testing, including hyperparameter tuning and cross-validation, the XGBoost classifier demonstrated similar performance in terms of the f1-score compared to the ANN model and other models we evaluated. The barplot of f1-scores for all the models illustrates that ANN and the XGBoost classifier achieved the highest f1-score.

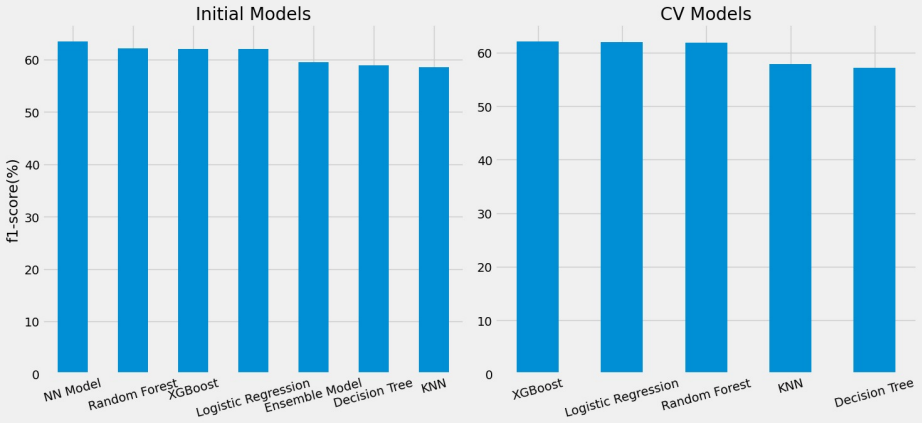

Fig. 2. F1 score (%) comparison

Data Preprocessing: Categorical features underwent one-hot encoding. Duplicate entries and rows containing missing values were removed. The dataset was partitioned into training and testing sets. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied.

Machine Learning Models:

- Algorithms: A variety of algorithms were evaluated, including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and XGBoost.
- Evaluation Metrics: Model performance was assessed using Accuracy, Precision, Recall, and F1-score.
- Model Selection and Hyperparameter Tuning: Randomized search with 3-fold stratified cross-validation was used for hyperparameter tuning, employing the F1-score as the optimization metric.
- Ensemble Method: A stacking classifier, utilizing the base models trained during hyperparameter tuning and a logistic regression final estimator, was constructed to potentially enhance predictive performance.

Deep Learning Model: We built a deep learning model for churn prediction using a feedforward neural network in PyTorch. This network had two hidden layers (64 and 128 units, respectively) with ReLU activation functions and a single output neuron for binary classification. The model was trained using the Adam optimizer and binary cross-entropy loss.

Results and Evaluation: To evaluate the deep learning model, we tracked the training progress by monitoring the loss and accuracy metrics across epochs. After training, we assessed the model's generalization performance on the unseen test dataset using metrics like test loss and accuracy. Finally, the trained model was used to predict churn labels for the test dataset, and these predictions were compared against the actual labels to understand the model's real-world efficacy.

2.4 Model Training

This section summarizes the performance of various machine learning models evaluated for customer churn prediction. Result can find in Table 2 below. Logistic Regression Model: The Logistic Regression Model demonstrated promising results for customer churn prediction, with good accuracy, precision, recall, and F1-score on both training and testing data. The model effectively identified churn cases, achieving a test accuracy of 76.65% and an F1-score of 65.11%. However, further analysis and fine-tuning are recommended to enhance performance and address potential class imbalance issues.

Decision Tree Classifier: The Decision Tree Classifier exhibited impressive performance on the training set, with near-perfect accuracy, precision, recall, and F1-score. However, it experienced a significant drop in performance on the test set, indicating potential overfitting. The model's interpretability and ability to capture non-linear relationships make it well-suited for churn prediction, but the observed discrepancy warrants further investigation to improve generalization.

Random Forest Classifier: The Random Forest Classifier, an ensemble learning method, showed remarkable training performance but suffered a noticeable drop in test performance. The model's complexity and overfitting to the training data may have contributed to this discrepancy. While Random Forests are inherently suitable for churn prediction due to their ability to handle complex relationships, the observed results suggest the need for hyperparameter tuning and addressing class imbalance for improved generalization.

XGBoost Classifier: XGBoost, a powerful gradient boosting algorithm, demonstrated good training performance but, similar to the Random Forest model, experienced a drop in test performance. Techniques to handle class imbalance were employed, yet the model may have still captured noise or irrelevant patterns, leading to suboptimal generalization. Further optimization and hyperparameter tuning are recommended to enhance XGBoost's performance on unseen data.

Conclusion: Overall, the models exhibited varying levels of performance, with some showing signs of overfitting and struggles with class imbalance. Further analysis, fine-tuning, and exploration of alternative modeling approaches are suggested to improve churn prediction accuracy and address the observed discrepancies between training and test sets. Regularization techniques, hyperparameter optimization, and addressing class imbalance through sampling techniques or cost-sensitive learning may prove beneficial in enhancing the models' performance and generalization capabilities.

2.5 Evaluation Measures

Evaluation measures for our project included Precision, Recall, and F1-score. Precision indicated the proportion of true positive predictions, while Recall measured the model's ability to correctly identify actual positive instances. The F1-score provided a balanced view, combining precision and recall into a single metric. By analyzing these measures, we gained insights into the performance of our models, particularly their effectiveness in identifying customer churn (class 1) and non-churn (class 0) instances. For example, the Logistic Regression model with SMOTE achieved a precision of 0.48 for churn instances, indicating a higher rate of false positives. In contrast, the XGBoost model without upsampling had a higher precision of 0.57 for churn, suggesting improved accuracy in identifying true churn cases. Additionally, the recall values offered information on the model's ability to correctly detect actual churn or non-churn instances. For instance, the LightGBM model had a recall of 0.89 for non-churn instances, indicating its strength in capturing most true negative cases. Overall, the F1-score provided a comprehensive view of model performance, balancing precision and recall. The varying F1-scores across different models and classes highlighted the trade-off between minimizing false positives and false negatives, helping us select the most suitable model for our churn prediction task.

3 Evaluation and Validation

3.1 Error Analysis

In this section, three experiments are conducted to analyze the performance of different machine learning models in predicting customer churn. The experiments aim to evaluate the models' ability to handle class imbalance and assess their suitability for the given dataset.

Experiment 1 (Logistic Regression with SMOTE and MinMax Scaling): Despite the implementation of SMOTE to address class imbalance, the recall and F1-score obtained from the logistic regression model are relatively low. The low recall indicates that the model is not effectively capturing all churn instances. This could lead to a significant number of false negatives, where actual

churn instances are incorrectly classified as non-churn. The suboptimal performance suggests that logistic regression might not be the most suitable algorithm for this dataset, even with techniques like SMOTE and MinMax scaling applied.

Experiment 2 (XGBoost without Upsampling or Scaling): XGBoost is evaluated on the original imbalanced dataset without up sampling or scaling techniques. By not addressing class imbalances, the model may be biased towards the majority class (non-churn), potentially leading to poor performance in correctly identifying churn instances.This experiment aims to determine if the inherent capabilities of XGBoost, such as its ability to handle complex datasets and learn non-linear relationships, can compensate for the imbalanced nature of the dataset and improve classification results compared to logistic regression.

Experiment 3 (LightGBM): LightGBM is known for its efficiency and performance on large datasets, making it a promising candidate for handling the imbalanced nature of the dataset.The error analysis for LightGBM focuses on observing if the model can effectively handle the class imbalance and provide improved classification results compared to logistic regression and XGBoost.By leveraging its gradient boosting framework and optimized tree-based learning algorithm, LightGBM may be able to capture intricate patterns and dependencies in the data, potentially leading to better performance in identifying churn instances.

3.2 Comparison with Kaggle notebooks

In our study, the highest weighted F1-score achieved was 74.9% using the XGBoost algorithm. The parameter `class_weight = balanced` significantly improved the F1-score, particularly for the positive class. We benchmarked our model against ten Kaggle notebooks that employed various models such as SVM, Logistic Regression (LR), XGBoost, AdaBoost, Random Forest (RF), and other advanced algorithms. Notably, one notebook using SMOTE and LR attained a 64.9% F1-score [2]. Another notebook achieved an impressive 83% F1-score using XGBoost with Chi-Squared Test for feature selection [3]. A standout notebook employed RF with the SMOTETomek sampling technique, achieving an F1-score of 84% [4].

4 Conclusion

The initial experiments evaluated three machine learning algorithms: logistic regression with SMOTE and MinMax scaling, XGBoost without upsampling or scaling, and LightGBM, for predicting customer churn in the telecommunications industry. Rigorous data preprocessing, including cleaning, visualization, feature engineering, and scaling, was performed to enhance data interpretability and

model training effectiveness. Each model exhibited strengths and weaknesses, contributing valuable insights into the challenges of churn prediction.

While the results were insightful, addressing class imbalance remains crucial for improving model performance. Future research should focus on exploring advanced techniques such as ensemble methods, feature engineering, and hyperparameter tuning to mitigate this issue. Experimenting with different classification algorithms and preprocessing techniques can provide valuable insights into identifying the most effective approach for churn prediction in the Telco dataset. Continuous monitoring and model updates with new data are essential to ensure relevance and adaptability to changing customer behavior and preferences.

By adopting a proactive approach to churn prediction and retention strategies, telecommunications companies can mitigate customer churn and foster long-term customer loyalty, thereby driving sustainable growth and profitability in the industry. Addressing class imbalance, exploring advanced techniques, and continuously updating models will be crucial in developing robust and accurate churn prediction models, enabling data-driven decision-making and customer retention strategies.

References

1. Zhang, T., Moro, S. and Ramos, R.F., 2022. A data-driven approach to improve customer churn prediction based on telecom customer segmentation. *Future Internet*, 14(3), p.94.
2. "Best techniques and metrics for Imbalanced Dataset." Accessed: May 16, 2024. [Online]. Available: https://kaggle.com/code/marcinrutecki/best-techniques-and-metrics-for-imbalanced-dataset
3. "Predicting Customer Churn with SMOTE | R." Accessed: May 16, 2024. [Online]. Available: https://kaggle.com/code/kellibelcher/predicting-customer-churn-with-smote-r
4. "Telco Churn: EDA|CV Score (85%+)| F1 Score (80%+)." Accessed: May 16, 2024. [Online]. Available: https://kaggle.com/code/tanmay111999/telco-churn-eda-cv-score-85-f1-score-80
5. "Telco Customer Churn." Accessed: May 17, 2024. [Online]. Available: https://www.kaggle.com/datasets/blastchar/telco-customer-churn

Table 1. Dataset structure

Attribute	**Attribute Explanation**
CustomerID	A unique identifier for each customer
Gender	Gender of the customer (e.g., Male, Female)
SeniorCitizen	Indicates if the customer is a senior citizen (1) or not (0)
Partner	Whether the customer has a partner or not (Yes, No)
Dependents	Whether the customer has dependents or not (Yes, No)
Tenure	Number of months the customer has stayed with the company
PhoneService	Whether the customer has a phone service or not (Yes, No)
MultipleLines	Whether the customer has multiple lines or not (Yes, No, No phone service)
InternetService	Type of internet service subscribed by the customer (DSL, Fiber optic, No)
OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)
OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)
DeviceProtection	Whether the customer has device protection or not (Yes, No, No internet service)
TechSupport	Whether the customer has tech support or not (Yes, No, No internet service)
StreamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)
StreamingMovies	Whether the customer has streaming movies or not (Yes, No, No internet service)
Contract	The type of contract the customer has (Month-to-month, One year, Two year)
PaperlessBilling	Whether the customer has paperless billing or not (Yes, No)
PaymentMethod	The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
MonthlyCharges	The amount charged to the customer monthly
TotalCharges	The total amount charged to the customer
Churn	The target variable indicates whether the customer churned (Yes) or not (No)

Table 2. Model Performance Metrics

Model	Data set	Accuracy	Precision	Recall	F1-score
XGBoost Classifier	**Train**	94.05%	90.96%	55.11%	88.44%
	Test	79.86%	63.86%	55.11%	59.16%
Random Forest Classifier	**Train**	99.8%	99.26%	100.0%	99.63%
	Test	80.93%	68.98%	50.81%	58.51%
Decision Tree Classifier	**Train**	99.8%	99.26%	100.0%	99.63%
	Test	73.17%	49.31%	48.12%	48.71%
Logistic Regression Model	**Train**	74.27%	50.86%	79.93%	62.16%
	Test	76.65%	53.87%	82.26%	65.11%