# MIDTERM REPORT: Predicting Customer Churn in the Telecom Industry Using Data Mining

**Team Members: Tajwar-Ul Hoque, Luis Escamilla, Isaiah Malcom**

## 1. Data

Our team is using the Telco Customer Churn dataset from Kaggle. This dataset includes 7,043 instances and 21 attributes that describe customer demographics, service subscriptions, billing information, and churn status. The target variable is Churn, which shows whether a customer left the company or not. The dataset contains both numerical and categorical variables, so we prepared it carefully for modeling.

In preprocessing, we first converted the TotalCharges column to numeric and filled missing values using the median to keep the data consistent. We dropped customerID since it has no predictive value. The Churn column was encoded as a binary variable, with "Yes" mapped to 1 and "No" mapped to 0. We then applied one-hot encoding to all categorical features to convert them into a numeric format suitable for machine learning models. Finally, we split the dataset into training (80%) and testing (20%) subsets, using stratified sampling to maintain the same churn proportion in both sets. These preprocessing steps were completed using pandas and scikit-learn in Python, ensuring the data is clean and ready for modeling.

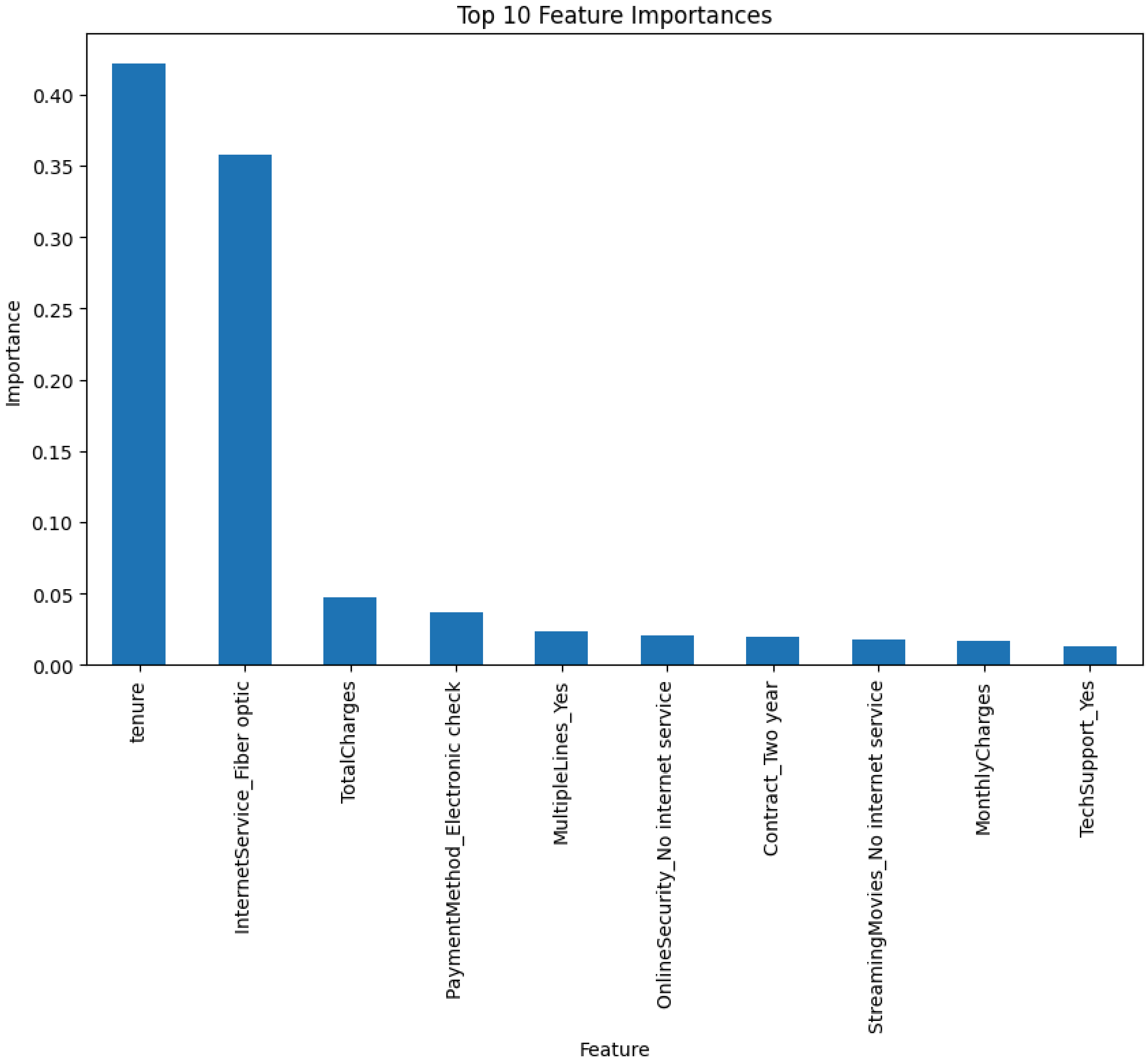
## 2. Data Mining Task

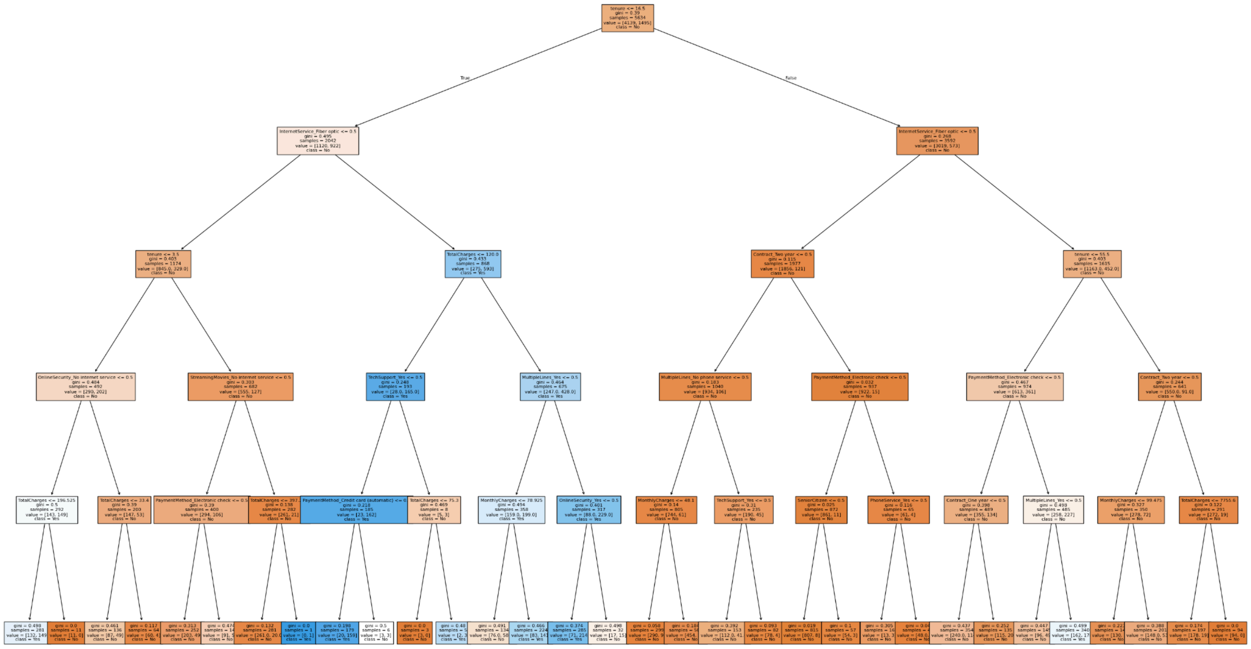
The main task for this project is classification. Our goal is to predict which customers are most likely to churn based on the available features. We are using supervised learning techniques for this purpose and plan to evaluate all models using accuracy, precision, recall, F1-score, and ROC-AUC. By building models that can accurately classify churn, we can identify at-risk customers and help telecom companies take proactive steps to retain them.

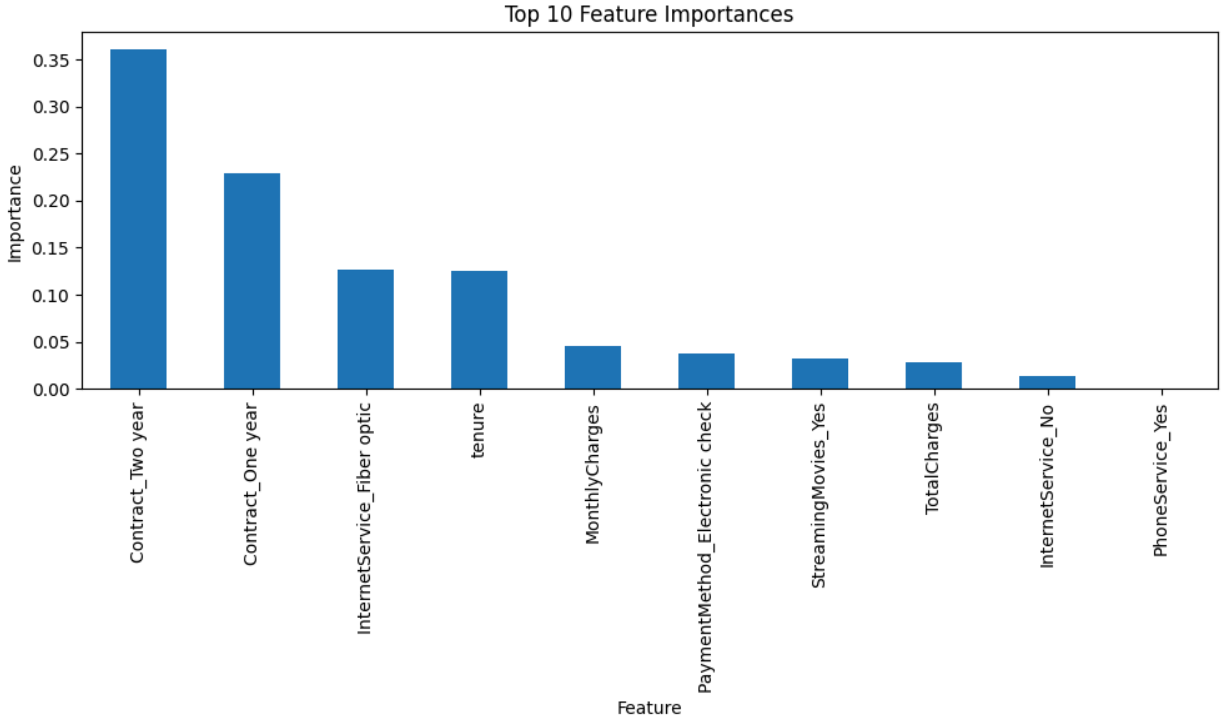
## 3. Progress

So far, our team has completed the first round of model training and evaluation using a Decision Tree Classifier as the baseline model. This model was trained using the Gini criterion with a maximum depth of 5. The Decision Tree achieved an accuracy of about 79% on the test set, showing solid baseline performance for this stage of the project. The results were further analyzed using classification reports and confusion matrices to verify consistency across precision and recall.

To better understand model behavior, we visualized the tree and saved it as tree.svg. We also generated a feature importance bar chart, saved as top_ten.svg, which highlights the top ten features contributing most to churn prediction. The most influential factors were Contract type, tenure, and MonthlyCharges, which align with expected business insights for telecom companies.


Top 10 Feature Importances
Importance
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
tenure
InternetService_Fiber optic
TotalCharges
PaymentMethod_Electronic check
MultipleLines_Yes
OnlineSecurity_No internet service
Contract_Two year
StreamingMovies_No internet service
MonthlyCharges
TechSupport_Yes
Feature


tenure <= 16.5 gini = 0.39 samples = 5634 value = [4139, 1495] class = No
True
False
InternetService_Fiber optic <= 0.5 gini = 0.495 samples = 2042 value = [1120, 922] class = No
InternetService_Fiber optic <= 0.5 gini = 0.268 samples = 3592 value = [3019, 573] class = No
tenure <= 3.5 gini = 0.403 samples = 1174 value = [845.0, 329.0] class = No
TotalCharges <= 120.0 gini = 0.433 samples = 868 value = [275, 593] class = Yes
Contract_Two year <= 0.5 gini = 0.115 samples = 1977 value = [1856, 121] class = No
tenure <= 55.5 gini = 0.403 samples = 1615 value = [1163.0, 452.0] class = No
OnlineSecurity_No internet service <= 0.5 gini = 0.484 samples = 492 value = [290, 202] class = No
StreamingMovies_No internet service <= 0.5 gini = 0.303 samples = 682 value = [555, 127] class = No
TechSupport_Yes <= 0.5 gini = 0.248 samples = 193 value = [28.0, 165.0] class = Yes
MultipleLines_Yes <= 0.5 gini = 0.464 samples = 675 value = [247.0, 428.0] class = Yes
MultipleLines_No phone service <= 0.5 gini = 0.183 samples = 1040 value = [934, 106] class = No
PaymentMethod_Electronic check <= 0.5 gini = 0.032 samples = 937 value = [922, 15] class = No
PaymentMethod_Electronic check <= 0.5 gini = 0.467 samples = 974 value = [613, 361] class = No
Contract_Two year <= 0.5 gini = 0.244 samples = 641 value = [550.0, 91.0] class = No
TotalCharges <= 196.525 gini = 0.5 samples = 292 value = [143, 149] class = Yes
TotalCharges <= 33.4 gini = 0.39 samples = 200 value = [147, 53] class = No
PaymentMethod_Electronic check <= 0.5 gini = 0.39 samples = 400 value = [294, 106] class = No
TotalCharges <= 397.2 gini = 0.138 samples = 282 value = [261, 21] class = No
PaymentMethod_Credit card (automatic) <= 0. gini = 0.218 samples = 185 value = [23, 162] class = Yes
TotalCharges <= 75.3 gini = 0.469 samples = 8 value = [5, 3] class = No
MonthlyCharges <= 78.925 gini = 0.494 samples = 358 value = [159.0, 199.0] class = Yes
OnlineSecurity_Yes <= 0.5 gini = 0.401 samples = 317 value = [88.0, 229.0] class = Yes
MonthlyCharges <= 48.1 gini = 0.14 samples = 805 value = [744, 61] class = No
TechSupport_Yes <= 0.5 gini = 0.31 samples = 235 value = [190, 45] class = No
SeniorCitizen <= 0.5 gini = 0.025 samples = 872 value = [861, 11] class = No
PhoneService_Yes <= 0.5 gini = 0.116 samples = 65 value = [61, 4] class = No
Contract_One year <= 0.5 gini = 0.398 samples = 489 value = [355, 134] class = No
MultipleLines_Yes <= 0.5 gini = 0.498 samples = 485 value = [258, 227] class = No
MonthlyCharges <= 99.475 gini = 0.327 samples = 350 value = [278, 72] class = No
TotalCharges <= 7755.6 gini = 0.122 samples = 291 value = [272, 19] class = No
gini = 0.498 samples = 281 value = [132, 149 class = Yes
gini = 0.0 samples = 11 value = [11, 0] class = No
gini = 0.461 samples = 136 value = [87, 49] class = No
gini = 0.117 samples = 64 value = [60, 4] class = No
gini = 0.313 samples = 252 value = [203, 49 class = No
gini = 0.474 samples = 14 value = [91, 5 class = No
gini = 0.132 samples = 281 value = [261.0, 20.0 class = No
gini = 0.0 samples = 1 value = [0, 1] class = Yes
gini = 0.198 samples = 179 value = [20, 159] class = Yes
gini = 0.5 samples = 6 value = [3, 3] class = No
gini = 0.0 samples = 3 value = [3, 0] class = No
gini = 0.48 samples = 5 value = [2, 3 class = Yes
gini = 0.491 samples = 134 value = [76.0, 58 class = No
gini = 0.466 samples = 224 value = [83, 141 class = Yes
gini = 0.374 samples = 285 value = [71, 214 class = Yes
gini = 0.498 samples = 32 value = [17, 15] class = No
gini = 0.058 samples = 299 value = [290, 9 class = No
gini = 0.184 samples = 5 value = [454, class = No
gini = 0.392 samples = 153 value = [112.0, 41. class = No
gini = 0.093 samples = 82 value = [78, 4] class = No
gini = 0.019 samples = 815 value = [807, 8] class = No
gini = 0.1 samples = 57 value = [54, 3] class = No
gini = 0.305 samples = 16 value = [13, 3 class = No
gini = 0.04 samples = 4 value = [48.0, class = No
gini = 0.437 samples = 354 value = [240.0, 11 class = No
gini = 0.252 samples = 135 value = [115, 20 class = No
gini = 0.447 samples = 14 value = [96, 49 class = No
gini = 0.499 samples = 340 value = [162, 17 class = Yes
gini = 0.22 samples = 14 value = [130, class = No
gini = 0.388 samples = 201 value = [148.0, 53 class = No
gini = 0.174 samples = 197 value = [178, 19] class = No
gini = 0.0 samples = 94 value = [94, 0] class = No



Our next steps include developing Logistic Regression, Random Forest, and XGBoost models , and potentially neural networks to compare their results against the baseline Decision Tree and determine which approach performs best.

## 4. Challenges

The main challenges we have faced involve handling class imbalance between churned and retained customers and balancing interpretability and performance. Decision Trees provide clear explanations but can struggle with overfitting if not tuned properly. The dataset is also moderately small, which limits the advantage of complex models. We plan to address these challenges by testing models with class weighting, performing cross-validation, and tuning hyperparameters for better generalization.

## 5. Schedule and Next Steps

At this point, we have completed data cleaning, preprocessing, and initial Decision Tree modeling. The upcoming phase of our project will focus on building and comparing additional models, including Logistic Regression, Random Forest, and XGBoost. We will also tune model hyperparameters using GridSearchCV to optimize performance. Once those models are evaluated, we will compare all results using multiple metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

We also plan to conduct a deeper feature importance analysis across all models and prepare visualizations to support our final report. The last stage will focus on compiling results, preparing visuals, and writing the final presentation and report.

## 6. Summary

Overall, our project is progressing according to the planned schedule. The preprocessing and baseline Decision Tree modeling steps were successful, giving us an initial accuracy of around 79%. The early results already highlight key factors affecting churn such as contract length, tenure, and monthly billing amounts. Moving forward, we aim to refine our predictive accuracy through advanced models and further analysis. This will help us provide stronger business and data-driven insights in the final stage of the project.

Link To The GitHub Repo:
https://github.com/Isaiah392/Data-Mining-Customer-Churn

**NOTE:** Please reach out to the team if you have any further question.