# CPS 844 Project Report

Mikael Syed
501112915
Mohamed Shrief
500981017
Group 30

# 1. Abstract

In this project, we used a Telco Customer Churn dataset from a telecommunications company. Our primary focus was to accurately predict whether a customer would churn by testing multiple data mining algorithms and examining the impact of feature selection on the data. Our initial step was to preprocess the dataset to address any missing values in the TotalCharges field; by doing so, we ensured the data was standardized into numerical values, fully compatible with our model and ready for use. Subsequently, we employed five algorithms that we believed would effectively aid us in determining precise predictions: Logistic Regression with L1 regularization, Decision Tree, Random Forest, Support Vector Machine and finally, Gradient Boosting, evaluating each with the use of 5 Fold Cross Validation. Performance and accuracy were examined using Accuracy, F1 Score and lastly ROC AUC. We have also used Recursive Feature Elimination (REF) to identify the top 10 most important features. Lastly, we compared our results with these selected features versus the full dataset, and the results and comparison will be discussed in depth throughout the project.

# 2 Dataset Introduction

The dataset used in this project is the Telco Customer Churn dataset, which we obtained from a public data repository on Kaggle. This dataset consists of 7,043 customer records from a telecommunications company. Key attributes in the dataset include:

1. **Demographic Features:**

   In the dataset, we identified several demographic data points, such as gender and

age. These attributes help us determine whether there are any churn patterns based on demographics

2. **Account and Service Details:**

   The dataset was rich in account and service details that helped us identify patterns based on tenure, including the types of phone services each client utilized along with any additional services like internet and online security. These attributes allowed us to understand how service usage could influence a customer's likelihood of churning.

3. **Billing Information:**

   Important billing attributes, such as Monthly Charges and Total Charges, are included. Although the Total Charges column initially had some inconsistencies, such as non-numeric values, we resolved these issues during our data pre-processing phase.

4. **Target Variable – Churn:**

   The churn column indicates whether a customer has left the service, which is our primary focus for prediction. This variable supports our modeling by offering a straightforward binary classification: churned or not churned.

# 3. Pre-processing and EDA

Initially, we had to clean our data from the dataset, so we began by importing it into our project. Then, we used commands such as df.info() and df.describe(include='all') to help us understand the data we were dealing with in terms of the number of records, data types, and basic statistics. After that, we ran the command df.isnull().sum() to check for missing values and found inconsistencies in TotalCharges. Consequently, we converted that column to a numeric format using the following code block:

```
df['TotalCharges'] = pd.to_numeric(df['TotalCharges'], errors='coerce')
print("\nMissing values in TotalCharges after conversion:",
df['TotalCharges'].isnull().sum())
df['TotalCharges'].fillna(0, inplace=True)
```

Then, categorical data were converted from “yes/no” to 1/0. Lastly, key numeric features such as tenure, MonthlyCharges, and TotalCharges were standardized using StandardScaler to align them on a similar scale with the following code block:

```
numeric_cols = ['tenure', 'MonthlyCharges', 'TotalCharges']
scaler = StandardScaler()
df[numeric_cols] = scaler.fit_transform(df[numeric_cols])
```

Lastly, for visualization, we created a bar chart to display the distribution of churn, histograms for the numeric features, and a correlation matrix heatmap to explore the relationships among the numeric features (Check the appendix for the graphs).

These steps helped us clean our data and prepare it for modeling. They also assisted us in choosing the appropriate methods and feature selections for the dataset.

# 4. Methodology

We have utilized various algorithms since they approach the dataset differently, providing us with the most reliable results regarding customer churn prediction.

## 4.1 Algorithm Selection

We have used 5 different algorithms to help us get the most accurate predictions:

1. **Logistic Regression (L1):** Our data includes a lot of linear relationship features with churn, such as MonthlyCharges. The simplicity which makes L1 regularization an ideal starting point

2. **Decision Tree:** Our data also included non-linear variables such as ContractType and InternetService. Since decision trees can capture non-linear relationships, we believed they would be the most effective choice for this aspect.
3. **Random Forest:** Since our data contained considerable noise, the random forest algorithm works best because each tree makes its own prediction, and then the random forest averages all of them to produce the most reliable answer.
4. **SVM (Linear Kernel):** Linear SVM was one of the best choices, as it provides clear coefficient values and ranks features by their importance, making it easier for us to select the most contributing ones for churn prediction.
5. **Gradient Boosting:** Since customer churn can depend on very small and varied patterns, we thought gradient boosting would be a great choice as it reviews previous models, catches mistakes, and results in better outcomes.

## 4.2 Feature Selection

For feature selection, we thought Recursive Feature Elimination (RFE) would be the best choice, as it reduces the number of features until it selects the top 10 most influential predictors for predicting churn. However, one challenge we faced was that it didn't work with SVM, because the RBF SVM version doesn't carry coef_ or feature_importance_. Therefore, we had to switch to the linear version.

## 4.3 Evaluation Metrics

For the Evaluation Metric, we picked the three key metrics:

1. **Accuracy:** This assesses the total percentage of accurate predictions.

2. **F1 Score:** As our dataset had imbalances, we used the F1 Score to prevent a class from dominating, as the F1 score balances precision and recall.
3. **ROC AUC:** This metric was a key option as it is good for binary classification tasks because it provides insight into the overall correctness of the model across different classes, specifically in our case, churn versus non-churn.

## 4.4 Cross-Validation

To ensure reliability throughout the model evaluations, we used 5-fold cross-validation, which splits the data into five sets. Each model is trained on four subsets and tested on one subset. This process is repeated five times, yielding an average result and providing a more reliable estimate of the model's performance.

# 5. Results

This section presents the experimental findings of our study, which evaluates and compares the performance of five supervised machine learning algorithms on the Telco Customer Churn dataset. The objective was to predict customer churn using classification models trained on two configurations of the dataset:

1. The full feature set, consisting of all pre-processed and encoded attributes
2. A reduced feature set, limited to the top 10 features selected using Recursive Feature Elimination (RFE).

All models were evaluated using 5-fold cross-validation to ensure consistency and reduce the risk of overfitting. The evaluation metrics are Accuracy, F1 Score, and ROC AUC were selected due to their relevance in binary classification tasks. In particular, F1 Score and ROC

AUC were included because they are more informative than accuracy when dealing with class imbalance, which is a common characteristic of churn prediction problems. These metrics collectively offer a balanced view of model performance across both majority and minority classes

## 5.1 Performance Evaluation Using the Full Feature Set

In the initial experiment, all of the five models—Logistic Regression (L1 penalty), Decision Tree, Random Forest, Support Vector Machine (SVM) using RBF kernel, and Gradient Boosting—were trained with the whole set of preprocessed features. These were a result of a lengthy preprocessing pipeline involving missing value management, conversion of categorical variables (by binary encoding and one-hot encoding), and numerical attribute feature scaling. This configuration aimed to offer a baseline performance against which the impact of feature selection would be compared later. The results obtained by means of this experiment are presented in Table 1.

Table 1: Performances of all 5 models with the full feature set

| Model | Accuracy | F1 Score | ROC AUC |
|---|---|---|---|
| Logistic Regression (L1) | 0.8036 | 0.5989 | 0.8453 |
| Decision Tree | 0.7298 | 0.4930 | 0.6559 |
| Random Forest | 0.7897 | 0.5477 | 0.8219 |

| | | | |
|---|---|---|---|
| SVM (Linear Kernel) | 0.7998 | 0.5871 | 0.8335 |
| Gradient Boosting | 0.8041 | 0.5892 | 0.8453 |

Table 1 shows that Logistic Regression and Gradient Boosting achieved the highest overall performance, while Decision Tree performed the weakest across all metrics. The other models demonstrated balanced results, with SVM (Linear) showing strong F1 and ROC AUC scores. A more detailed analysis is provided in Section 8.

## 5.2 Performance Evaluation Using Feature Selection

To reduce dimensionality and enhance model efficiency, Recursive Feature Elimination (RFE) was applied individually to each classifier to identify the top 10 most informative features. RFE is a wrapper-based feature selection method that recursively fits a model, ranks features based on their importance, and eliminates the least significant ones in successive iterations. To ensure compatibility with this method, all selected models were required to support either the coef_ or feature_importances_ attribute, which are essential for evaluating feature relevance during the elimination process. Table 2 displays the outcomes of this experiment.

Table 2: Performance of all 5 models with top 10 features selected using RFE

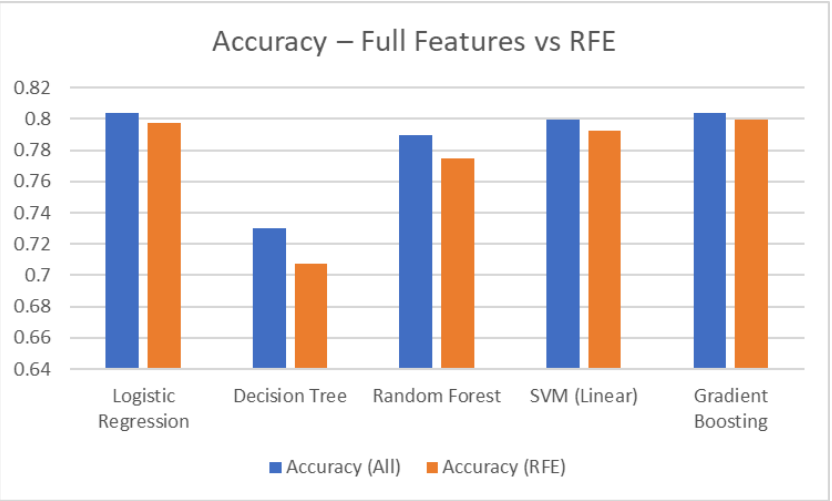
| Model | Accuracy | F1 Score | ROC AUC |
|---|---|---|---|
| Logistic Regression (L1) | 0.7972 | 0.5832 | 0.8428 |
| Decision Tree | 0.7077 | 0.4595 | 0.6330 |

| | | | |
|---|---|---|---|
| Random Forest | 0.7750 | 0.5352 | 0.8110 |
| SVM (Linear Kernel) | 0.7923 | 0.5665 | 0.8274 |
| Gradient Boosting | 0.7997 | 0.5838 | 0.8431 |

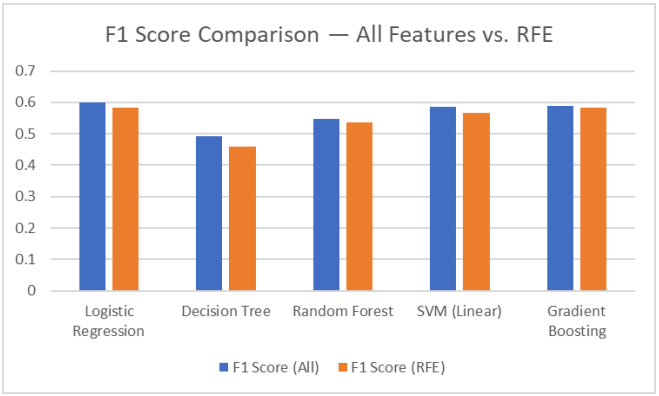
Table 2 shows the performance of all models using the top 10 features selected by RFE. Logistic Regression and Gradient Boosting continued to perform well, with minimal drops in accuracy and ROC AUC. SVM (Linear) remained consistent across all metrics, while Random Forest showed a slight decline. The Decision Tree had the most significant performance drop, especially in ROC AUC. A more detailed analysis of these differences is discussed in Section 8.

## 5.3 Visual Comparison of Performance

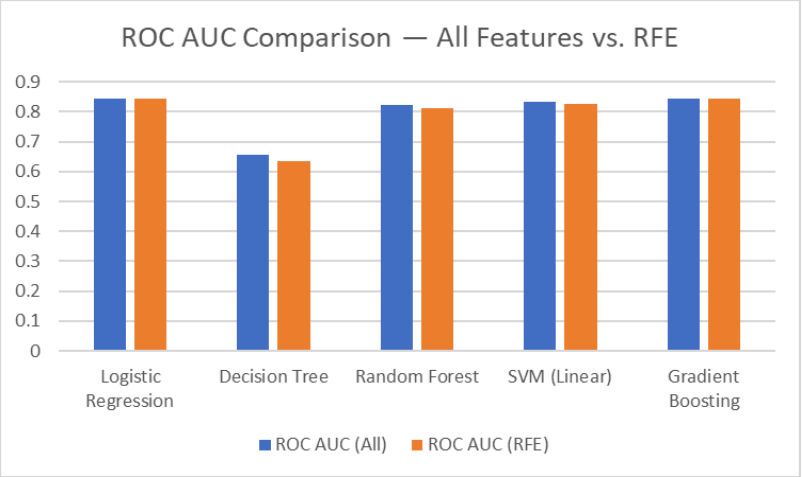
Visual comparisons were conducted to complement the numerical evaluation and provide a clearer illustration of the effects of feature selection across models. The following bar charts display the Accuracy, F1 Score, and ROC AUC for each model, trained on both the full feature set and the RFE-selected top 10 features. Presenting these metrics side by side highlights subtle differences in model performance and offers a more intuitive understanding of how dimensionality reduction influenced each classifier’s effectiveness.




In the above graph, the accuracy comparison of all models trained on the full feature set versus the top 10 features selected using RFE. While all models show a slight decrease in accuracy after feature selection, the performance remained relatively stable, with Logistic Regression and Gradient Boosting maintaining the highest accuracy overall.




For the above graph, F1 Score comparison between models trained on the full feature set and those using the RFE-selected features. The results show minimal changes across most models, with Logistic Regression, Gradient Boosting, and SVM (Linear) maintaining strong F1 scores. The Decision Tree model saw the most noticeable drop, indicating sensitivity to feature reduction.




According to the chart above, ROC AUC comparison for all models using the full feature set and the RFE-selected features. Across all models, ROC AUC scores remained largely consistent, with only slight decreases observed post-RFE. Logistic Regression and Gradient Boosting maintained the highest discriminatory power, while Decision Tree exhibited the largest drop, suggesting reduced class-separation capability with fewer features.

# 6. Discussion

## 6.1 Comparison of Models

Out of all the five classifiers evaluated, the Gradient Boosting and Logistic Regression achieved the highest overall performances across all the three values measured, which were Accuracy, F1 Score and ROC AUC, and that in both cases, the one with all features and with the top 10 features according to RFE. Gradient Boosting had the top F1 score and had a high ROC AUC score as well, which matched Logistic Regression as well, and this is referring to both the scenarios. This shows that the Gradient Boosting model has the ability to handle more complex relationships in the data and can deal with the class imbalance between churn and non-churn customers. In relation to Logistic Regression was competitive in the ROC AUC score, expressing it has strong capabilities of classification of churners and

non-churners in the model. Both these models had similar F1 scores in both cases, which convey that they have a moderate balance of identifying churners, as they do identify many but not all. Furthermore, the Support Vector Machine has consistent performances in all categories, especially having an ROC AUC score of 0.8335 when all features are used and 0.8271 when feature selection is applied, which exhibits that the data was linearly separable and that SVM's margin-maximizing approach was effective. On the other hand, the Random Forest model delivers a moderate performance, showing balanced results all around, even though it lagged behind Gradient Boosting and SVM in terms of F1 Score and ROC AUC. This conveys that it was dependable but was slightly less effective at separating the two classes and at pointing out the churners compared to the other top-performing models. Lastly, the Decision Tree classifier showed the weakest performance across the board, particularly in ROC AUC. This suggests that the model overfitted the training data and wasn't able to perform well on new, unseen data. The detectable drop in performance after feature selection portrays that single decision trees are sensitive to the number of features and can struggle when there's less information to split on.

## 6.2 Effect of Feature Selection

Using Recursive Feature Elimination (RFE) had the slightest impact on the overall model performance. In most metrics, minimal decreases were noticed as the model was reduced to taking in the top 10 features and the drop did not have a huge impact on the effectiveness of the model. For example, with Gradient Boosting and SVM, the F1 and ROC AUC scores remained virtually the same, which showed that the RFE successfully identified and retained the most informative features.  The trade-off of this compromise was simpler models, faster

training time, and more interpretability, which are beneficial in practical applications where model efficiency and interpretability are important.

## 6.3 Interpretation of Feature Importance

Feature importance analysis revealed certain variables to always be selected by high-performing models. The Contract type, tenure, MonthlyCharges, and PaperlessBilling were most frequently selected by RFE and had high rankings in the Decision Tree feature importance and Random Forest importance values. These are intuitively significant features as they represent longer-term customer behavior, billing, and payment tendencies that are all good indicators of the likelihood of churn.

# 7. Conclusion

In this study, five machine learning models were evaluated on the Telco Customer Churn dataset to assess their performance with and without feature selection using Recursive Feature Elimination (RFE). Gradient Boosting and Logistic Regression emerged as the top-performing models across all metrics, maintaining strong accuracy and classification performance even when reduced to the top 10 features. This suggests that feature selection can simplify models and reduce computational cost without significantly sacrificing predictive power. These findings have practical implications for businesses aiming to deploy churn prediction systems efficiently with fewer inputs, enabling faster, targeted decision-making.

For future work, exploring additional feature selection methods, such as SHAP values or Lasso, testing other advanced models like XGBoost, and tuning hyperparameters may further enhance model performance and robustness in real-world applications.

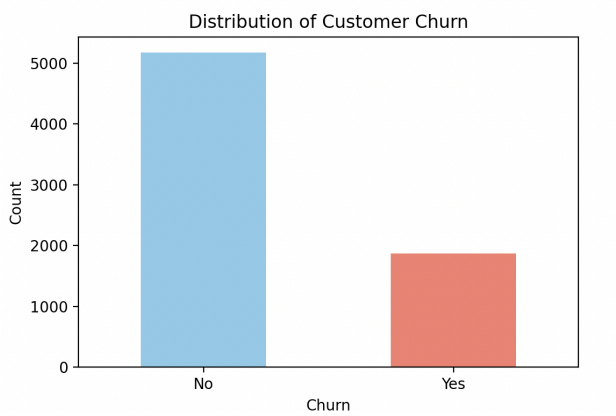
# 7. Appendix and Citation

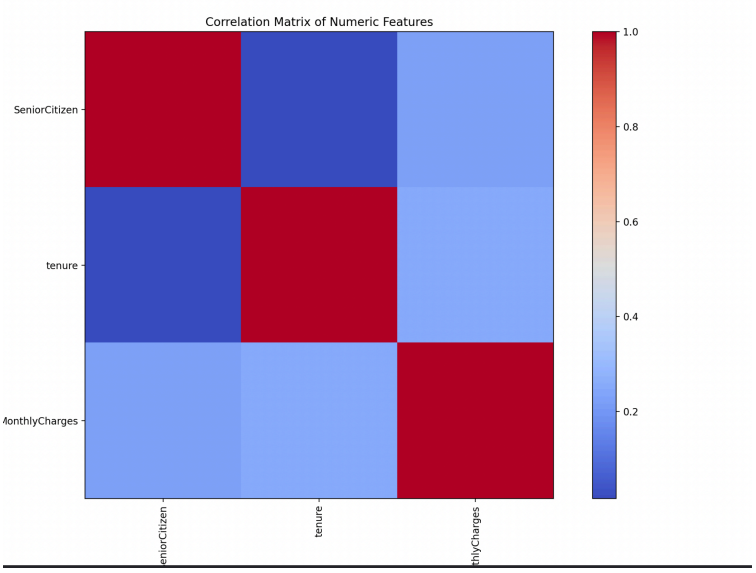
## 7.1 Citation:

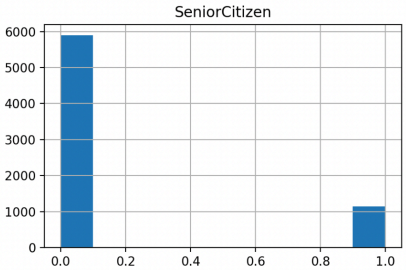
BlastChar. (2018, February 23). *Telco customer churn*. Kaggle. https://www.kaggle.com/blastchar/telco-customer-churn

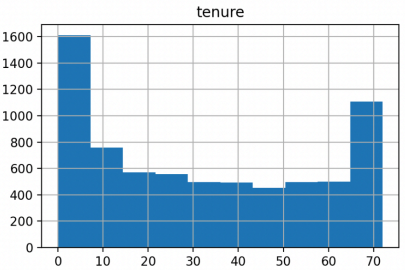
## 7.2 Appendix:

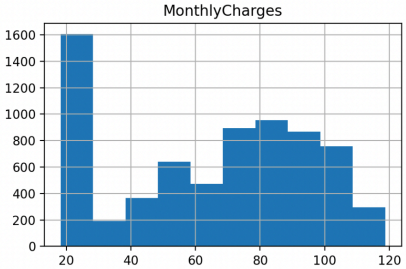
Graphs from EDA and Pre-processing step:


Distribution of Customer Churn
Count
5000
4000
3000
2000
1000
0
No
Yes
Churn


Correlation Matrix of Numeric Features
SeniorCitizen
tenure
MonthlyCharges
1.0
0.8
0.6
0.4
0.2


SeniorCitizen
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0


tenure
1600
1400
1200
1000
800
600
400
200
0
0
10
20
30
40
50
60
70


MonthlyCharges
1600
1400
1200
1000
800
600
400
200
0
20
40
60
80
100
120