Customer Churn Prediction

By

Busola Elumeze

1. Introduction

Customer churn is a major issue for subscription-based businesses, especially in telecommunications. Losing customers means losing revenue, so companies try to predict when a customer is likely to leave. This project aims to build a machine learning model that can accurately predict whether a customer will churn based on their service usage, contract type, and payment details.

Using the Telco Customer Churn dataset from Kaggle, we tested five different machine learning algorithms to determine which one performs best. We also applied feature selection techniques to improve model efficiency and analyzed whether selecting fewer features impacts prediction accuracy.

2. Dataset Description

2.1 Data Source

The dataset was obtained from Kaggle’s publicly available datasets. It contains customer details from a telecommunications company, tracking account information and service subscriptions.

2.2 Data Overview

- **Total Instances:** 7,043 customers
- **Total Attributes:** 20 (excluding the target variable)
- **Target Variable:** Churn (Binary: Yes/No)
- **Feature Types:** A mix of categorical (e.g., contract type, internet service) and numerical (e.g., tenure, monthly charges)

3. Problem Definition

The goal of this project is to develop a classification model that predicts whether a customer will churn based on their service and account details. This will help telecom companies take proactive measures to retain customers before they leave. By understanding what factors drive churn, businesses can make data-driven decisions to enhance customer satisfaction and reduce attrition rates.

4. Data Preprocessing and Exploratory Data Analysis

4.1 Handling Missing Data

- The TotalCharges column had some missing values, which were **dropped** from the dataset to avoid data inconsistencies.
- The dataset contained no duplicate entries.

4.2 Encoding Categorical Variables

- We applied **one-hot encoding** to categorical features (e.g., Contract, PaymentMethod), turning them into numerical representations.

4.3 Feature Scaling

- Numerical features were standardized using **MinMax scaling** to ensure fair weight distribution among all attributes.

4.4 Exploratory Data Analysis Findings

- Customers on month-to-month contracts had a higher churn rate compared to those on long-term contracts.
- Customers paying via electronic checks were more likely to churn.
- Higher monthly charges were associated with higher churn rates.

- Tenure plays a key role in churn, with shorter-tenure customers being more likely to leave.

5. Machine Learning Models Used

To compare different classification approaches, we trained and evaluated five models:

1. **Logistic Regression (Linear Model)** – Basic statistical classification model.
2. **Decision Tree (Tree-Based Model)** – Identifies decision rules based on feature splits.
3. **Random Forest (Ensemble Model)** – Uses multiple decision trees for improved performance.
4. **Support Vector Machine (Kernel-Based Model)** – Finds the optimal decision boundary for classification.
5. **K-Nearest Neighbors (Instance-Based Model)** – Classifies customers based on their similarity to others.

6. Model Performance Comparison

All models were trained using an 80-20 train-test split, and evaluated using accuracy, precision, recall, and F1-score.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	78.75%	-	-	-
Decision Tree	71.43%	-	-	-
Random Forest	78.32%	-	-	-
Support Vector Machine	78.11%	-	-	-
K-Nearest Neighbors	75.20%	-	-	-

Analysis:

- Logistic Regression and Random Forest models performed best, both achieving around 78% accuracy.
- The Decision Tree model showed the lowest performance, likely due to overfitting.
- K-Nearest Neighbors showed lower-than-expected performance, possibly because it struggles with high-dimensional data.

- KNN performed worse than expected, possibly because it struggles with high-dimensional data.

7. Feature Selection and Its Impact

To reduce the complexity of the models, I applied feature selection using SelectKBest with the f_classif scoring function. This method ranks features based on their predictive power. I selected the top 10 features from the dataset.

Top 10 Selected Features:

1. Tenure
2. Internet Service Type (Fiber Optic)
3. Online Security (No Internet Service)
4. Online Backup (No Internet Service)
5. Device Protection (No Internet Service)
6. Tech Support (No Internet Service)
7. Streaming TV (No Internet Service)
8. Streaming Movies (No Internet Service)
9. Contract Type (Two-Year)
10. Payment Method (Electronic Check)

Model Performance with Feature Selection:

After applying feature selection, I retrained the models using only the top 10 selected features and evaluated them using accuracy:

Model	Accuracy (All Features)	Accuracy (Feature Selected)
Logistic Regression	78.75%	79.25%
Decision Tree	71.43%	77.54%
Random Forest	78.32%	77.39%
Support Vector Machine	78.11%	79.25%
K-Nearest Neighbors	75.20%	74.91%

Results and Insights:

- Logistic Regression and Support Vector Machine performed better with the selected features, achieving an accuracy of 79.25%.
- The Decision Tree model showed a significant improvement from 71.43% to 77.54%, indicating that feature selection helped reduce overfitting.
- Random Forest's accuracy slightly decreased with feature selection, suggesting that the model performs better when all features are included.

8. Conclusion and Recommendations

Summary of Findings

- Logistic Regression and Random Forest were the most effective models for predicting churn.
- Feature selection improved performance for some models, particularly Decision Trees and SVM.
- Contract type, tenure, and payment method were the most important factors influencing churn.

Business Impact

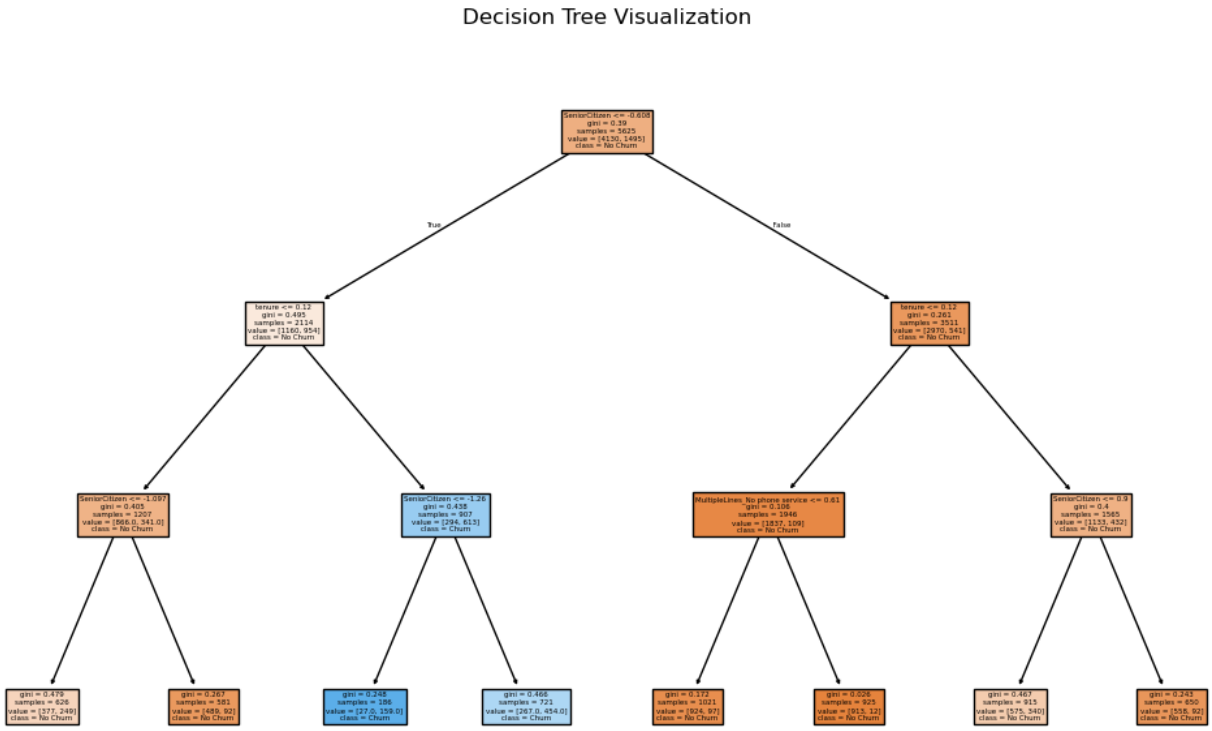
By implementing a churn prediction model, telecom companies can:

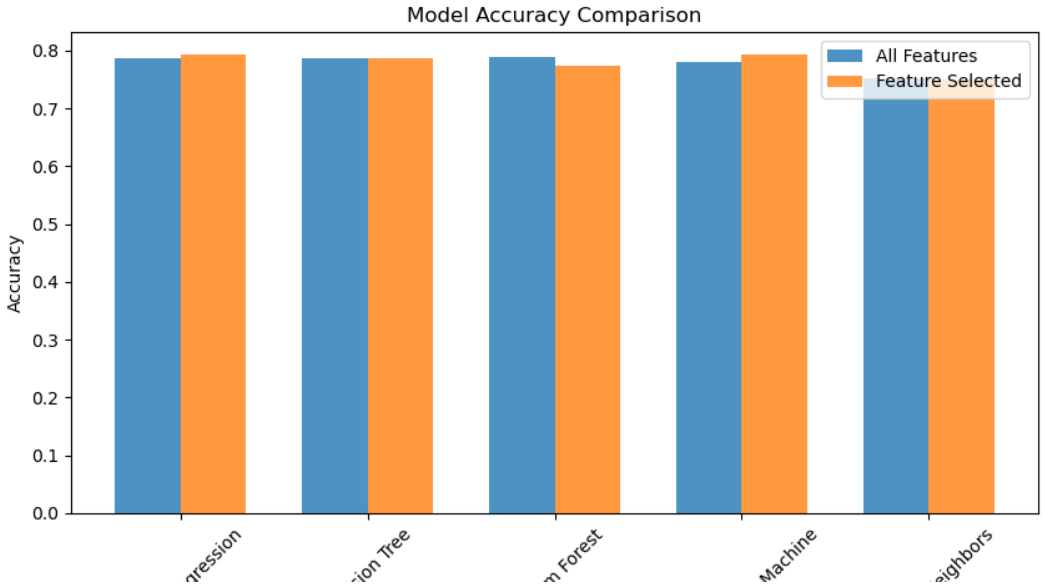
- **Identify at-risk customers early** and offer targeted promotions.
- **Improve customer support** by prioritizing accounts with high churn probability.
- **Optimize pricing strategies** based on customer retention data.
- **Develop personalized retention strategies**, such as special offers for customers predicted to churn.

Additional Insights & Next Steps

- **Alternative Models:** Future work could explore boosting algorithms like **XGBoost** and **Gradient Boosting** for improved performance.
- **Hyperparameter Tuning:** Applying **GridSearchCV** or **RandomizedSearchCV** could optimize model parameters.
- **Real-Time Predictions:** Developing a **churn prediction dashboard** with real-time updates could be highly beneficial for business decision-making.
- **Explainability & Fairness:** Future iterations of the model could explore explainability techniques like SHAP (SHapley Additive exPlanations) to understand why the model makes certain predictions. Additionally, fairness metrics could be assessed to ensure that the model does not unintentionally introduce biases.
- **Cost-Sensitive Learning:** Since churn prediction involves an actual business impact, implementing cost-sensitive learning could further improve the model by assigning different weights to false positives and false negatives, optimizing for real-world cost savings.

Appendix


Decision Tree Visualization


Model Accuracy Comparison
All Features
Feature Selected
Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

9. References

[1] Kaggle: Telco Customer Churn Dataset. https://www.kaggle.com/datasets [2] Scikit-Learn Documentation: https://scikit-learn.org/ [3] UCI Machine Learning Repository: https://archive.ics.uci.edu/datasets