Customer Attrition Modeling

Abhay Kothari

Data Set & Project Information

Telco Customer Churn Dataset

The data set includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

Project

Analyze all relevant customer information and create predictive models to predict if customers will churn or not churn.

Data Cleaning

Missing Values

Found 11 missing values

Duplicates

126 duplicate values were found and dropped

Data Preprocessing

Features & Target

Features : All features
Target : Churn

Scaling

StandardScaler was used

Dataset Splitting

Dataset was split 75-25
(Train-Test)

Statistical Tests

For all categorical features, chi-square tests of independence were ran to assess the dependence between each categorical variable and Churn.

Dependent Features:

1. **Partner** and churn are dependent.
2. **Dependents** and churn are dependent.
3. **Multiple lines** and churn are dependent.
4. **Internet service** and churn are dependent.
5. **Online security** and churn are dependent.
6. **Online backup** and churn are dependent.
7. **Device protection** and churn are dependent.
8. **Tech support** and churn are dependent.
9. **Streaming TV** and churn are dependent.
10. **Streaming movies** and churn are dependent.
11. **Contract and churn** are dependent.
12. **Paperless billing** and churn are dependent.

Independent Features:

1. **Gender** and churn are independent.
2. **Phone service** and churn are independent.

Logistic Regression

This model was used to determine whether a given customer would churn or not churn: “Yes” or “No”.

Best subset selection for this model revealed best combination of features was all of them.

Using the training x and y values,

After fitting the logistic regression model (logreg), it achieved :

	Recall Score
Mean Cross-Val	0.846
Test	0.835
Train	0.848

Decision Tree Classifier

This model was used to understand the underlying patterns and interactions between the input features.

Best subset selection for this model revealed best combination of features was all of them.

Using the training x and y values,

After fitting the decision tree model (dt) it achieved :

	Recall Score
Mean Cross-Val	0.802
Test	0.779
Train	0.998

Random Forest Classifier

This model improves generalization by leveraging the diversity of decision trees on random subsets of data and features.

Best subset selection revealed best combination of features was all of them.

After fitting the random forest model (rf) it achieved :

	Recall Score
Mean Cross-Val	0.85
Test	0.843
Train	1.0

XGBOOST Classifier

This model helped identify non-linear relationships between the input features to improve classification results for churn.

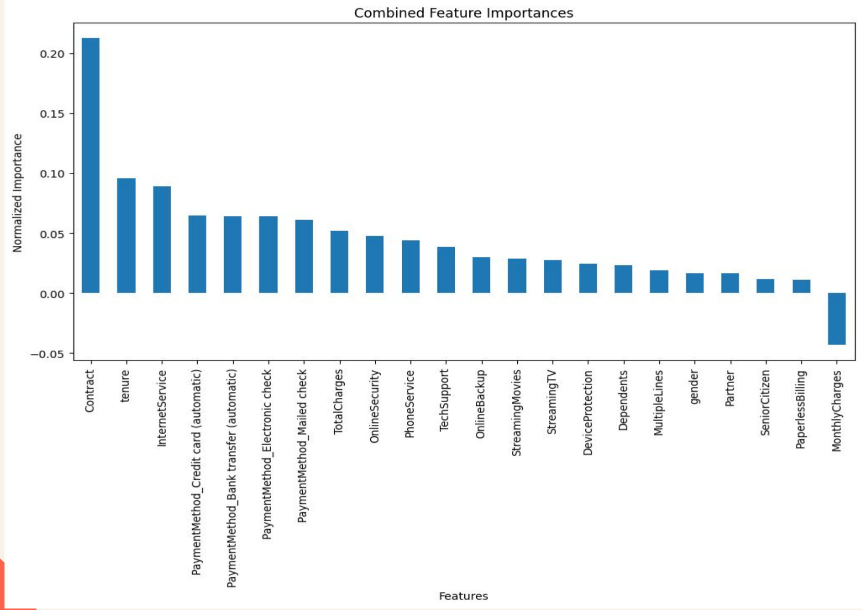
Best subset selection revealed best combination of features was all of them.

After fitting the XGBoost model (xgb) it achieved :

	Recall Score
Mean Cross-Val	0.846
Test	0.996
Train	0.98

Feature Importance

This analysis aims to understand the relative importance of each feature in classifying churn. It provides insights into which features contribute the most to the model's predictions. The combined feature importance was calculated by normalizing the feature importance values from four different classification models and then averaging them.



Top 4 Predictor Variables	Value
Contract	0.212
tenure	0.096
Internet Service	0.089
PaymentMethod - Credit Card	0.065

Final Observations

Best Model

- XGBoost is the most reliable model for predicting churn, with high and consistent recall scores across both training and test data.

Contract Type

- Customers with month-to-month contracts are significantly more likely to churn, indicating the importance of longer-term contracts to reduce churn rates.

Service Subscriptions

- Subscriptions to InternetService without additional services (OnlineSecurity, TechSupport, OnlineBackup, or DeviceProtection) increase the likelihood of churn.

Customer Demographics

- Customers without a partner or dependents have a higher churn rate.

Payment Methods

- Customers using electronic check for payments are more prone to churn, implying that marketing alternative payment methods might reduce churn.

Feature Importance

- The most influential features in predicting churn are contract type, tenure, and InternetService. These features should be prioritized when trying to mitigate churn.