# **TELECOM CHURN ANALYSIS**

**1. Problem Statement:**

In the telecommunications sector, a significant issue arises from customer churn, which refers to customers discontinuing their use of a company's services. To combat this, it's pivotal for companies to pinpoint customers inclined to unsubscribe and to take measures to retain them. Leveraging Machine Learning models, we aim to predict potential churners based on various parameters, including customer usage patterns, payment history, and demographic information.

**2. Primary Purpose:**

Our primary objective is to employ predictive models to ascertain whether a telecom customer will churn. These models either predict a churn probability for each customer or classify them into churners and non-churners. Such forecasts can be invaluable for identifying customers on the brink of leaving and devising proactive strategies to retain them.

**3. Data Source:**

We have used open-source data set called Telco-customer-churn.csv from <https://www.kaggle.com/> that is [Telco Customer Churn](https://www.kaggle.com/bhartiprasad17/customer-churn-prediction/data). We can see that the data contains 21 columns and 7044 rows.

* Churn Status: Customers who left within the previous month.
* Subscribed Services: Includes phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
* Account Information: Details such as customer tenure, contract type, payment method, billing details, and charges.
* Demographics: Information like gender, age bracket, and family details (partners and dependents).

**4. Implementation:**

* Libraries Used: sklearn, Matplotlib, pandas, seaborn, and NumPy.
* Data Loading: The dataset was fetched from Google Drive, followed by datatype conversion for the TotalCharges column in the telco\_df dataframe. We also inspected the dataset for null or missing values, identifying 11 missing values for Total Charges.

**Data Loading:**

We have loaded the dataset from the google drive using the following code.

A screenshot of a computer

Description automatically generated

Then we Converted the ‘TotalCharges’ to a numerical data type in the ‘telco\_df’ dataframe.

We have also checked for any null or missing values.

A screenshot of a computer program

Description automatically generated

After looking at the output, we can say that there are 11 missing values for Total Charges. Let us replace and remove these 11 rows from our data set.

**5. Data Cleaning:**

The first step in the data cleaning process was to remove any missing values in the dataset.

Convert all categorical variables into dummy variables and convert churn predictor variable to binary numeric value.

A screenshot of a computer

Description automatically generated

**6. Exploratory Data Analysis:**

In this step we have performed EDA i.e, Exploratory Data Analysis on the data, where we get valuable insights out of it.

Firstly, we performed Correlation of “Churn” (outcome or dependent variable) with other variables in ‘telco\_df\_dummies’.

A screenshot of a computer code

Description automatically generated

A graph with different colored lines

Description automatically generated

This correlation graph gives us,

**Key Observations:**

* Month-to-month contracts, lack of online security, and tech support tend to drive churn.
* Longer customer tenure and two-year contracts correlate with reduced churn.
* Surprisingly, services like online security, TV streaming, online backup, tech support, etc., without an internet connection, appear to deter churn.

**7. Visualization Insights**

|  |  |  |  |
| --- | --- | --- | --- |
| A graph with blue and orange squares  Description automatically generated | | | A blue circle with orange triangle and black text  Description automatically generated |
| **Demographic Analysis**: Our dataset is evenly split between male and female customers. Gender appears to have no discernible influence on the churn rate. A minority (16%) of our clientele are senior citizens, indicating a predominantly younger demographic. | | | |
| A graph of a number of people with numbers  Description automatically generated with medium confidence | | | A graph with a number of people  Description automatically generated with medium confidence |
| Half of the customers have partners, with only 30% having dependents. Among those with partners, there's an equal split between those with and without dependents. Contrastingly, 80% of single customers have no dependents. | | | |
| A graph of customers by their certain amount  Description automatically generated with medium confidence | | | A graph of blue and orange bars  Description automatically generated |
| A graph with blue lines  Description automatically generated | | | |
| There's a higher frequency of customers with short-term tenures. Churn rate seems particularly elevated among month-to-month contract subscribers, while two-year contract holders display much lower churn rates. Notably, most monthly contracts have a lifespan of 1-2 months, whereas two-year contracts typically last around 70 months, indicating greater customer loyalty with longer-term contracts. | | | |
| A group of blue and white bars  Description automatically generated | | | |
| The visualizations shed light on the distribution and popularity of different services among customers. | | | |
| A graph of a graph of a chart of a graph  Description automatically generated with medium confidence | A graph of a number of churn rate  Description automatically generated | | |
| The data provides insight into the connection between monthly and total charges among the customers. | | | |
| A graph of a number of different colored rectangular objects  Description automatically generated with medium confidence | | A screenshot of a graph  Description automatically generated | |
| Only 26% of the customers in our dataset churned, while a majority, 74%, stayed loyal to the service. It's vital to recognize the skewness in our dataset, as it can impact our predictive modeling, potentially leading to a rise in false negatives | | | |
| A graph of a number of people  Description automatically generated with medium confidence | | A graph of a distribution of charge  Description automatically generated | |
| A graph of distribution of charge  Description automatically generated | | | |
| Loyal customers have a noticeable trend; they tend to stay longer with the telecom company. Customers on a month-to-month contract exhibit a significantly higher churn rate. | | | |

**8. Machine Learning Models Employed:**

1. Logistic Regression:

* Overview: A statistical technique predominantly used for binary classification problems.
* Application: Considering our target variable is a binary outcome (Churn/No Churn), we utilized Logistic Regression, achieving an accuracy of 79%.
* Performance: The ROC curve presented an AUC (Area Under Curve) of 0.83.

2. Decision Tree Model:

* Overview: Decision trees map decisions in a branching structure, ideal for both classification and regression problems.
* Application: Using this approach, we achieved a model accuracy of 72%.
* Performance: The ROC curve showed an AUC of 0.65, with a true positive rate of 0.81, false positive rate of 0.52, and precision of 0.81.

3. Random Forest Model:

* Overview: An ensemble method, this classifier employs multiple decision trees, enhancing predictive accuracy.
* Application: For predicting Churn/No Churn, the model displayed an accuracy of 78%.
* Performance: Its ROC curve depicted an AUC of 0.81. The model had a true positive rate of 0.89, false positive rate of 0.53, and precision of 0.82.

4. K-Nearest Neighbors (KNN):

* Overview: KNN, a supervised ML algorithm, predicts values based on how closely new data points match those in the training set.
* Application: Using KNN, we achieved an accuracy of 77%.
* Performance: The ROC curve resulted in an AUC of 0.74. The model's true positive rate was 0.88, false positive rate was 0.55, and precision was 0.82.

5. Naïve Bayes Classifier:

* Overview: This supervised machine learning algorithm calculates the likelihood of data points belonging to specific classes based on Bayes' theorem. It assumes feature independence given the class label.
* Application: The model's accuracy was determined to be 68%.
* Performance: The ROC curve showed an AUC of 0.68. Performance metrics include a true positive rate of 0.62, false positive rate of 0.16, and precision of 0.92.

## A screenshot of a computer program Description automatically generatedA screenshot of a computer program Description automatically generated

A screen shot of a computer program

Description automatically generatedA graph and chart with numbers

Description automatically generatedA screenshot of a computer program

Description automatically generated

A blue squares with white text

Description automatically generatedA screenshot of a computer program

Description automatically generated

A graph with a line and a blue dotted line

Description automatically generated

A screenshot of a computer program

Description automatically generated

A diagram of a confusion matrix

Description automatically generated A graph with a line and a blue line

Description automatically generated

A screenshot of a computer program

Description automatically generated

A diagram of blue squares

Description automatically generated A graph with a line and a blue line

Description automatically generated

**A screenshot of a computer program

Description automatically generated**

A graph of different models

Description automatically generated

A computer screen shot of a program

Description automatically generated

A graph of a number of different colored lines

Description automatically generated

Upon comparing the five predictive models (Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbors, and Naïve Bayes Classifier), the following conclusions were drawn:

In terms of accuracy, Logistic Regression proved to be the superior model, while the Naïve Bayes Classifier yielded the lowest accuracy.

Our ROC curve comparisons further validated the supremacy of Logistic Regression over other models in predicting customer churn.

The predictive models have not only been successfully constructed but have also been compared to draw definitive conclusions about their efficacy.

**9. Dimension Reduction and Clustering:**

* Now I'd like to discuss the insights and conclusions drawn from our recent data analysis, particularly focusing on dimension reduction, clustering, association rules, and neural networking.
* Our journey started with the goal to simplify our vast dataset. Utilizing PCA, we condensed the dimensionality of our data, capturing 95% of the variance with just 15 principal components. This not only reduced computational load but also made visualization and analysis more intuitive.

A screenshot of a computer

Description automatically generated

A graph of different colored dots

Description automatically generated

**INTERPRETATION FOR DIMENSION REDUCTION, CLUSTERING & PCA:**

**Red Cluster Insights:**

* Contains younger customers (only 17.68% are senior citizens).
* Almost half have partners, and slightly over a quarter have dependents.
* Most of them use phone services, and a significant portion uses multiple lines.
* A substantial number uses the internet, but only about 27% have online security.
* The average monthly charge is roughly $64.63, and the average total charge is $2158.76.
* The churn rate is relatively high at 27.11%.

**Green Cluster Insights:**

* Similarly, this cluster contains younger customers (17.55% are senior citizens).
* A slightly higher proportion have partners compared to the red cluster, and also a bit higher proportion have dependents.
* The service usage profile is similar to the red cluster but with a lower percentage of internet users.
* The average monthly and total charges are $62.82 and $2299.45, respectively.
* The churn rate is the lowest among the three at 22.70%.

**Blue Cluster Insights:**

* Has the least proportion of senior citizens among the three clusters.
* Contains the highest proportion of individuals with partners and dependents.
* Most of them use phone services, and a significant portion uses multiple lines. The internet usage percentage is also the highest among the three.
* The average monthly and total charges are the highest among the three clusters.
* The churn rate sits in the middle at 24.32%.

**From our observations:**

* The Green cluster seems to be the most stable with the lowest churn rate.
* The Red cluster experiences the highest churn rate, suggesting we might need interventions to retain these customers.
* The Blue cluster, despite spending the most, also has a relatively high churn rate.

A graph with blue dots

Description automatically generated

**10. PCA Analysis**:

* The first component captures ~99.99% of variance.
* By the third component, nearly 100% variance is explained.
* Data likely has correlated features; dimensionality reduction is viable.

A graph with a line

Description automatically generated

**K-Means Clustering**:

* Major inertia drops from 1 to 2 clusters.
* Significant reduction again from 2 to 3 clusters.
* Diminishing reductions in inertia after 3 clusters.
* Optimal cluster count: 2 or 3, based on inertia.

**11. ASSOCIATION RULES ANALYSIS & INTERPRETATION:**

**Association Rule Analysis Overview**

**- Goal**: Identify relationships between customer behavior (features) using transaction data.

**- Method**: Applied the Apriori algorithm to identify frequent item sets.

**- Metrics Used**:

* + **Support**: Probability of occurrence of the itemset.
  + **Confidence**: Likelihood of Y happening, given X occurs.
  + **Lift**: Increase in the ratio of the sale of Y when X is sold.

A screenshot of a computer

Description automatically generated

**Key Findings (Sampled Data)**

- Based on a 10% sample of the data:

* + **74.9%** of the users with **Fiber Optic** internet service prefer a **Month-to-month** contract.
  + **80.3%** of users who opt for **Electronic Check** payments are on a **Month-to-month** contract.
  + Customers combining **Fiber Optic** with **Electronic Check** payments are **1.47 times** more likely to opt for a **Month-to-month** contract.
  + **72.3%** of users with both **Electronic Check** payments and **Month-to-month** contract use **Fiber Optic** as their internet service.

**12. NEURAL NETWORKS’ INTERPRETATION:**

**Neural Network Model**

* **Objective**: To predict customer churn based on various features.
* **Model Details**:
  + A simple 3-layer neural network.
  + 32 neurons in the first two layers with ReLU activation.
  + A single neuron in the output layer with sigmoid activation (binary classification).
  + Adam optimizer with a learning rate of 0.001 was used.
* **Results**:
  + The model achieved ~98.81% accuracy on the training set by the 10th epoch.
  + Validation accuracy remained consistent at 72.41% from the 3rd to the 10th epoch.

**Observation**: Although the model performs well on the training set, the consistent validation accuracy suggests potential overfitting. Consider implementing regularization techniques or adjusting the architecture.

A screenshot of a computer program

Description automatically generated

**Neural Network Classifier Performance**

* **Objective**: To validate the trained model's performance.
* **Results**:
  + **Overall Accuracy**: 76%
  + **Precision** (of predicting Churn = Yes): 81%
  + **Recall** (of actual Churn = Yes instances): 89%
  + **F1-Score** (harmonic mean of precision and recall for Churn = Yes): 85%
* **Confusion Matrix**:
  + True Negatives (Correctly predicted 'No Churn'): 186
  + False Positives (Incorrectly predicted 'Churn'): 24
  + False Negatives (Incorrectly predicted 'No Churn'): 44
  + True Positives (Correctly predicted 'Churn'): 28

**Observation**: The model is more accurate at predicting customers who don't churn (recall of 89% for 'No Churn'). However, it struggles somewhat with identifying those who do (recall of 39% for 'Churn'). This is reflected in the F1-score, which gives a balanced measure of the model's performance on both classes.

**13. FINDINGS & RECOMMENDATIONS:**

**Major Findings from Data Analysis:**

1. **Customer Segmentation:** K-Means clustering identified distinct customer groups, enabling targeted marketing.
2. **Churn Association:** Fiber optic users show a high association with churn, indicating possible service dissatisfaction.
3. **Service Affinities:** Association rules highlight prevalent service combinations, suggesting popular package preferences.
4. **Churn Prediction:** Our neural network effectively predicts customer churn with 76% accuracy.
5. **Model Insights:** The precision and recall metrics emphasize successful identification of non-churners, though there's room for improvement in identifying actual churners.
6. **Potential Improvements:** Continuous feedback from churn-prone segments can provide real-time insights for immediate action.

**Recommendations Based on Data Insights:**

1. **Tailor Marketing:** Design campaigns based on customer clusters from KMeans.
2. **Enhance Fiber Optic Service:** Address higher churn among fiber optic users through service improvements or feature additions.
3. **Optimize Service Bundles:** Use association rules to bundle popular service combinations, boosting sales.
4. **Targeted Retention:** Utilize the neural network's churn predictions to offer loyalty incentives to high-risk customers.
5. **Implement Feedback:** Create mechanisms for feedback, especially from potential churners, to understand and address concerns.
6. **Refine Prediction Model:** Further develop the neural network model for increased accuracy and actionable insights.

**Conclusion:**

We have applied different predictive models to the data dataset telco-customer-churn.csv and obtained evaluation criteria like accuracy, precision, ROC curve. Based on those we have determined that the Logistic Regression is the best model out of all. Then did some further analysis and obtained some insights to render in the future analysis.