Telecom Customer Churn Analysis Report

Abstract

Customer attrition poses a significant threat to telecommunication companies, impacting top-line and long-term growth. The objective of this project was to identify prospective customers who would churn using historical customer data with machine learning models. Steps of analysis involved data cleaning, feature generation, correlation analysis, and class imbalance handling by SMOTE. Various models of classification—Random Forest, XGBoost, Decision Tree, and Balanced Bagging—were tried using precision, recall, and F1-score as criteria with a special emphasis on recall to minimize false negatives. Key predictors such as contract type, tenure, and service features were derived. The findings were presented through an interactive dashboard, and strategic recommendations were provided to maximize customer retention. This data-driven approach allows proactive prevention of churn and helps in better decision-making for telecom operators.

Problem Statement

Telecom operators are still facing high rates of customer churn, with customers dropping services due to dissatisfaction with service quality, being offered better deals by competitors, or having unmet expectations. This kind of churn not only reduces revenue but also increases the cost of acquiring new consumers and disrupts long-term planning.

Traditional retention strategies typically reactive in nature fail to expose the underlying causes of churn. By actively analyzing customer demographics, service usage, billing preferences, and behavior, companies can anticipate churn and take proactive measures to retain customers. This project will create a predictive model that identifies customers who are likely to churn using historical customer data.

Aim and Objective

Aim:

To identify potential churn customers using binary classification models by analyzing demographic, billing, and service-related features and addressing data imbalance for more accurate predictions.

Objectives:

- Perform data preprocessing and cleaning
- Conduct exploratory data analysis (EDA) and correlation analysis

- Perform feature engineering for modeling
- Handle imbalanced data using SMOTE
- Build and evaluate classification models
- Develop a dashboard to communicate insights
- Recommend business actions based on analysis

Data Collection

The dataset is derived from a telecom company and includes the following attributes:

- **Demographics**: gender, Senior Citizen, Partner, Dependents
- **Account info**: tenure, Contract, Paperless Billing, Payment Method
- **Services**: Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies
- **Financials**: Monthly Charges, Total Charges
- **Target Variable**: Churn

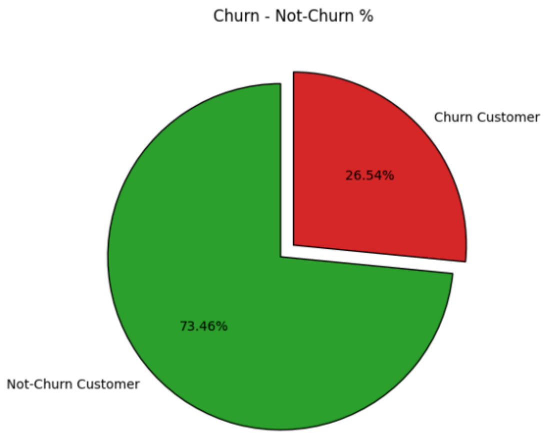
Data Cleaning

Key steps taken:

- **Type conversion**: Total Charges was found to be of type object due to empty strings and was converted to float after handling missing values.
- **Missing Values**: Handled missing or blank entries in Total Charges.
- **Duplicate Records**: Ensured there are no duplicate records especially in “customerID” column.
- **Columns Creation:** Created few columns which help to discover more generalize insights through EDA and dashboards.
- Also, used IQR method to detect outliers which influence predictions in modelling.

Exploratory Data Analysis (EDA)

For EDA, initially, columns are divided into 2 parts: Numerical and Categorical. Furthermore, the categorical features are breakdown into 3 categories: Demographic, Service-Based and Contract & payment behavior.



Target feature distribution:

As, it can be seen that the churn customer rate is 26.54%, which means that quarter number of customers are leaving every year.

Also, this indicates that dataset in bias towards non-churn customers so it would be better to handle this class imbalance to avoid building bias model.

Categorical Feature Analysis

Demographic Insights:

- **Gender**: Churn rate for male and female are almost similar.
- **Senior Citizen**: The number of Senior Citizen customers is relatively low. However, among Senior Citizens, approximately 40% experienced churn, accounting for 476 out of 1142 Senior Citizen customers.
- **Partner & Dependents**: Customers without partners or dependents were more likely to churn.

Service-Based Insights:

- **Internet Service**: Fiber optic users had the highest churn.
- **Online Security & Backup**: Customers lacking these services showed higher churn.
- **Device Protection and Tech Support:** Customers lacking these services showed higher churn.
- **Streaming services**: Those with streaming services showed moderate churn risk.
- **Multiple Lines** and **Phone Service** presence slightly reduced churn likelihood.

Contract & Payment Behaviour:

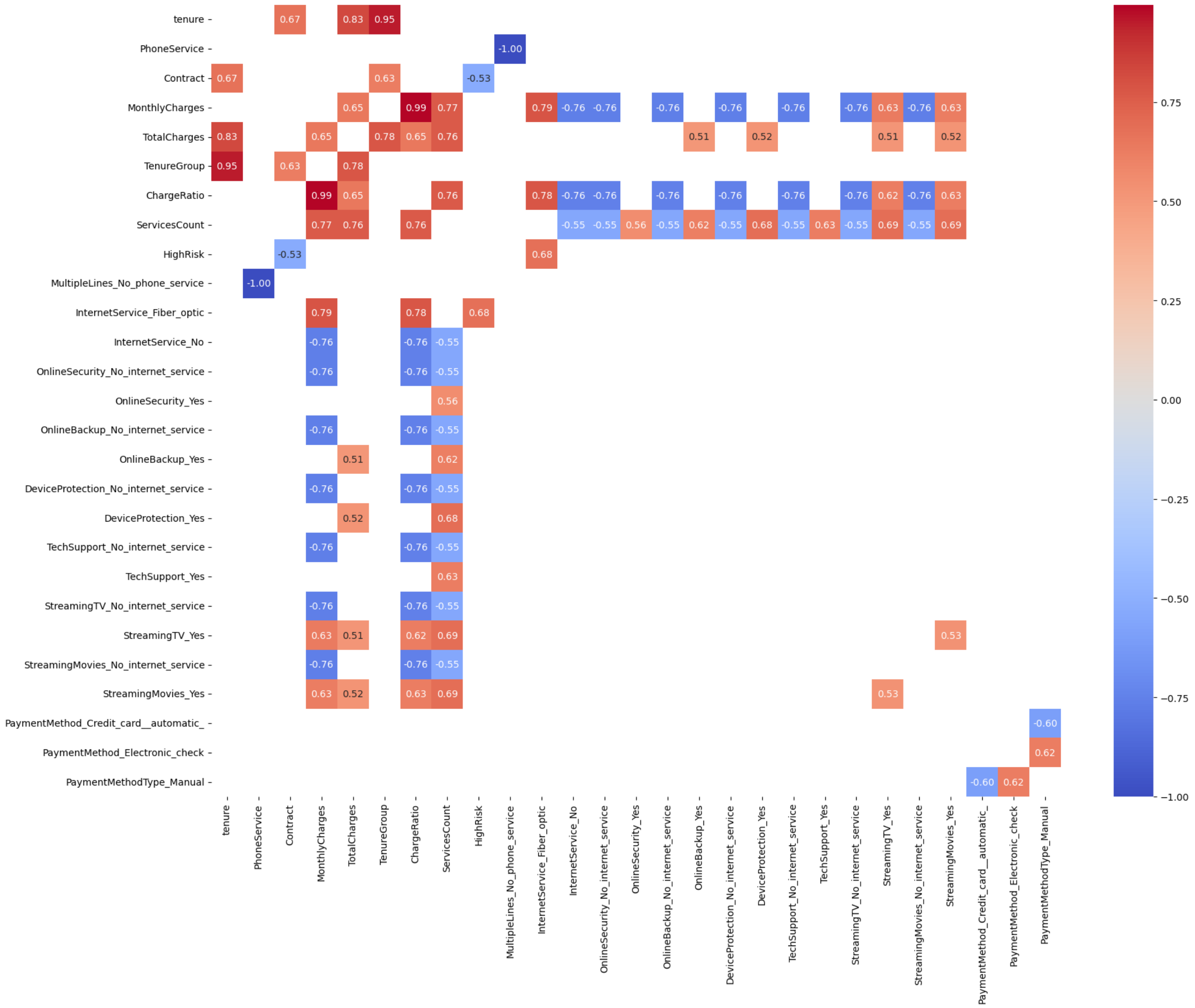
- **Month-to-month contracts**: Significantly higher churn rates than 1- or 2-year contracts.
- **Paperless Billing**: Associated with higher churn.
- **Electronic Check Payment**: Customers using this method showed the highest churn compared to other methods.

Numerical Feature Analysis

- Customers are more likely to churn due to the presence of high monthly charges.
- Based on the total charges and Tenure, it is observed that as customer spent more time their total charges are increased, and churn probability tend to decrease.
- **Monthly Charge Threshold:** Customers tend to decide to cancel their subscriptions when MonthlyCharges reach 70 and above.
- **Tenure Relation:** Customer with lower tenure (0-20) are likely to churn, while higher tenure customers are only at risk of churn if monthly charge surpass the threshold value.

Feature Engineering

- Drop irrelevant column like “customerID” which does not contribute to modelling.
- Used different types of the encoding according to features like one-hot encoding, label encoding.
- Create correlations matrix to avoid multicollinearity.



Correlations matrix Insights

- **Tenure, Contract, and Charges**: Longer tenure strongly correlates with higher total charges and longer contracts. Customers with more services or fiber optic internet tend to pay higher monthly charges.
- **Risk and Internet Absence**: Longer contracts reduce customer risk. Customers without internet or specific services (like security or streaming) pay significantly less.
- **Churn Patterns**: Churn is higher among high-risk customers, fiber optic users, and those with short tenure. Longer contracts and loyal customers are less likely to churn.
- **Service Usage and Revenue**: More services lead to higher total charges. Online backup and streaming service usage also correlates with increased revenue.

Key Insights on Churn:

- **Churn Drivers**: High-risk customers, fiber optic users, and electronic check payers are more likely to leave.
- **Retention Factors**: Long-term contracts, loyal customers (longer tenure), and those without extra services (e.g., streaming) stay longer.
- **Actions**: Target high-risk users, improve fiber service value, push for contract renewals, and simplify payment processes.

Before modelling, numerical features are scaled through standard scaler method.

Feature Selection:

Filter the features which are less contributing in predicting the churn variable, I applied the “ANOVA F-value.” Then, based on the f-score, features are removed whose score was less than 100.

Model Building & Evaluation:

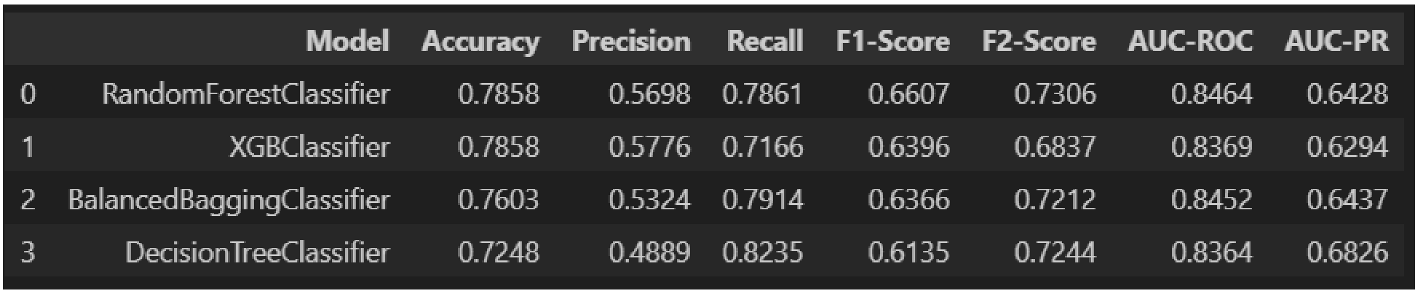
Before model building, I splatted the data into 3 part training (70%), validation (10%) and testing (20%).

Through EDA, we discovered that dataset was imbalanced by 1:4 ratio. To solve this problem, I applied the SMOTE technique where I did over-sampling (increasing record of minor class) to balance the number of churn and non-churn record in training set.

Here, I used the models which help to classify the correct class. So, I used the model like Random Forest, XGBoost Classifier, Decision Tree, and Balanced Bagging Classifier. Initially, I added base parameter values which provides general overview to discover best performance model.

I look for the metrics like precision, recall, f1-score and f2-score along with the accuracy to evaluate these models. I focus on recall as false negatives (missed churners) are costlier than false positives.

Model Comparison



	Model	Accuracy	Precision	Recall	F1-Score	F2-Score	AUC-ROC	AUC-PR
0	RandomForestClassifier	0.7858	0.5698	0.7861	0.6607	0.7306	0.8464	0.6428
1	XGBClassifier	0.7858	0.5776	0.7166	0.6396	0.6837	0.8369	0.6294
2	BalancedBaggingClassifier	0.7603	0.5324	0.7914	0.6366	0.7212	0.8452	0.6437
3	DecisionTreeClassifier	0.7248	0.4889	0.8235	0.6135	0.7244	0.8364	0.6826

The **RandomForestClassifier** is the best model for balancing both precision and recall, with a recall of **0.7861** and precision of **0.5698**. Its **F1-Score (0.6607)** reflects a strong overall performance, making it effective for distinguishing churn cases while managing false positives. Additionally, with a **AUC-ROC (0.8464)** and **AUC-PR (0.6428)**, it excels in handling imbalanced data, ensuring reliable predictions and a good trade-off between identifying churners and reducing false positives.

Furthermore, I removed the columns that cause the multicollinearity based on the correlation matrix. After that, I tune the **RandomForestClassifier** using Grid search to find out the best parameters value, then I train that model with training and validation dataset to evaluate with test dataset. Finally, I get the model with the 75% accuracy and 71% recall.

At the end, with the help of Flask framework, I connect front-end with backend, where model can predict the output based on the user inputs.

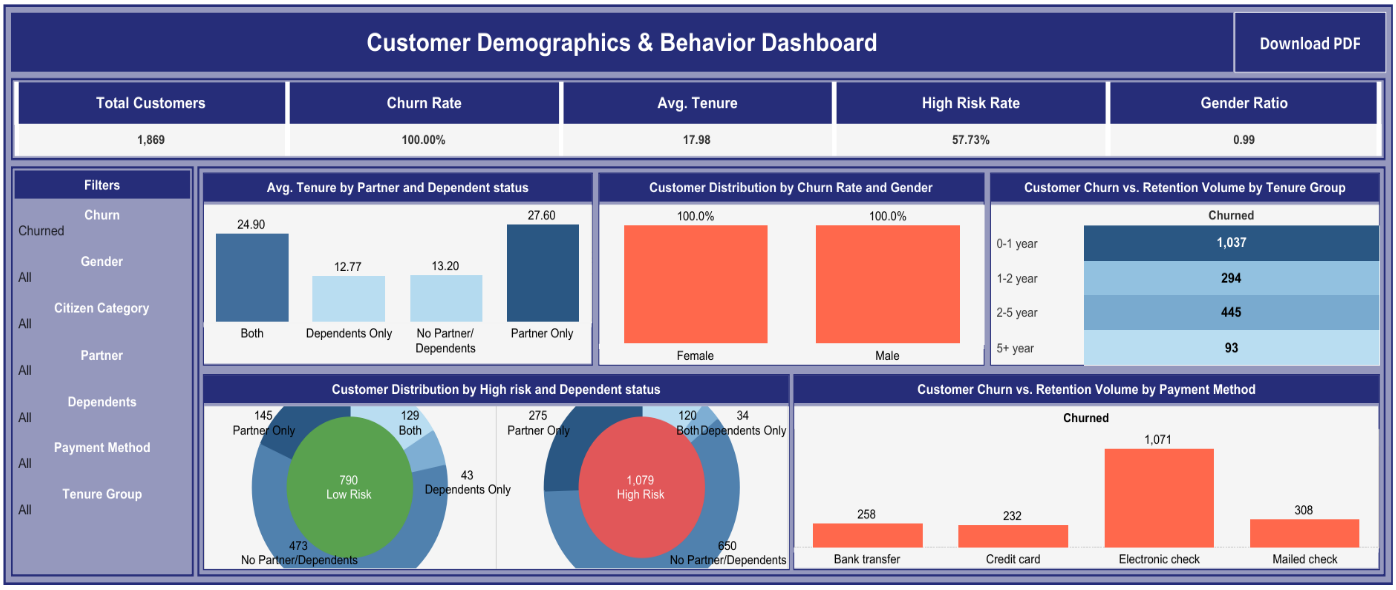
Dashboard Insights:

To effectively communicate insights from the churn analysis and support business decision-making, I developed an interactive dashboard. The dashboard serves as a **visual storytelling tool**, allowing stakeholders to explore customer churn patterns, segment risks, and service-related trends without needing to analyze raw data directly. It bridges the gap between technical modeling and strategic action by making data-driven findings more accessible and actionable.

For clarity and usability, the dashboard is divided into **two key sections**:

1. **Customer Demographics & Behavior Dashboard**
2. **Customer Service & Billing Dashboard**

Demographics & Behavior Dashboard



Customer Demographics & Behavior Dashboard

Download PDF

Total Customers	Churn Rate	Avg. Tenure	High Risk Rate	Gender Ratio
1,869	100.00%	17.98	57.73%	0.99

Filters

Churn: Churned

Gender: All

Citizen Category: All

Partner: All

Dependents: All

Payment Method: All

Tenure Group: All

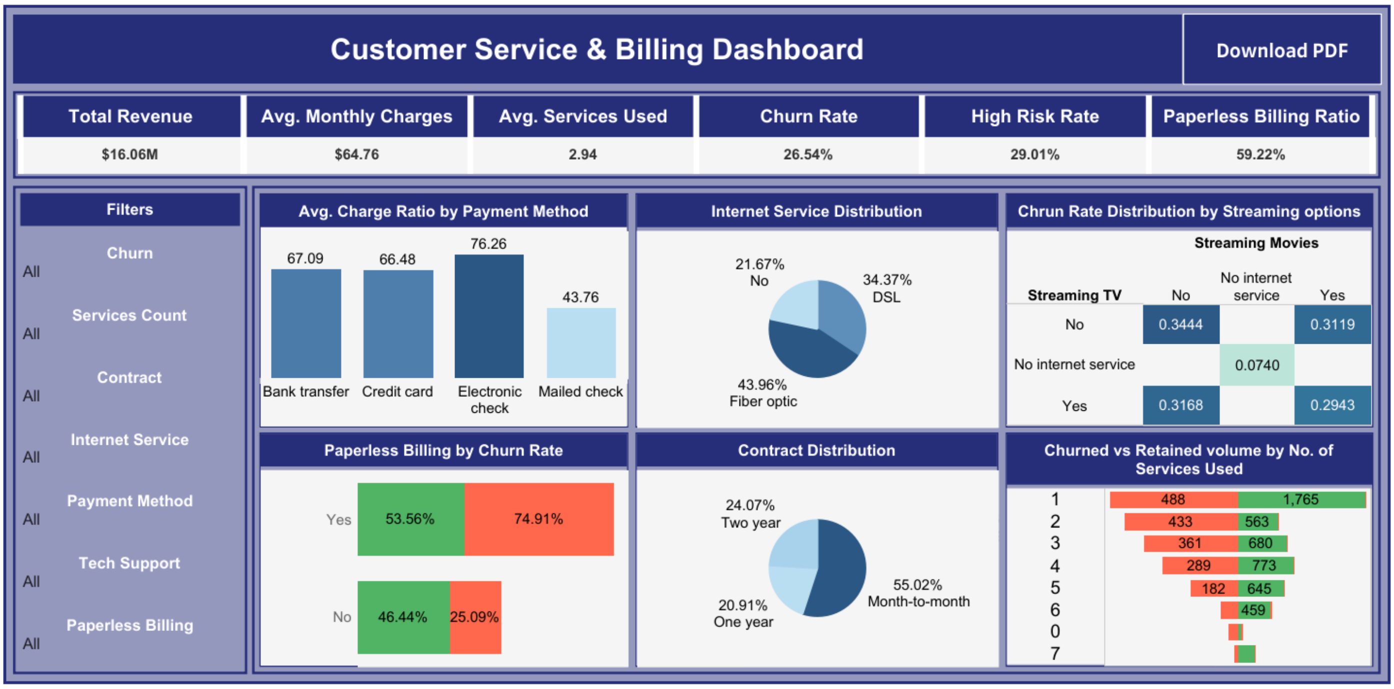
Insights:

- Gender does not significantly influence churn behavior. The churn rate is nearly the same across both genders.
- The **newer the customer, the higher the churn rate**. Customers with under 1 year of tenure have nearly **7x the churn rate** of long-term customers.
- Customers with a partner (especially with both partner & dependents) stay significantly longer.
- Customers without family ties are more likely to be high risk.
- **Electronic check users have an extremely high churn rate**, far above other payment methods.

Recommendations:

- Implement welcome bonuses, loyalty points, or educational campaigns early on.
- Personalize offers to engage this high-risk group — e.g., individual-focused plans.
- Encourage payment method change or investigate experience issues.
- Consider retention packages or “refer a partner/dependent” schemes.

Customer Service and Billing Dashboard



Customer Service & Billing Dashboard

Total Revenue	Avg. Monthly Charges	Avg. Services Used	Churn Rate	High Risk Rate	Paperless Billing Ratio
$16.06M	$64.76	2.94	26.54%	29.01%	59.22%

Filters

Churn: All

Services Count: All

Contract: All

Internet Service: All

Payment Method: All

Tech Support: All

Paperless Billing: All

Chrun Rate Distribution by Streaming options

Streaming TV	Streaming Movies: No	No internet service	Yes
No	0.3444		0.3119
No internet service		0.0740	
Yes	0.3168		0.2943

Insights:

- Churn is **highest for customers with fewer services (1–2)**, Balance for those who used **4-5 services** and less for more service users.
- **Month-to-month contracts (55.02%)** dominate, which are more prone to churn.
- **Fiber optic** is the most used (43.96%) followed by **DSL**, also the churn rate is higher in **Fiber optic.**
- Customers without internet but using **no streaming services** show **lowest churn rate (0.0740)**.
- **Electronic check** customers have the **highest average charges ($76.26)**, hence people are more likely to leave who use this payment method.
- customers who **did not opt** for paperless billing are **less likely to churn.**

Recommendation:

- Encourage bundling of services with discounts or value-added perks.
- Incentivize long-term contracts with exclusive offers.
- Cross-sell internet services to customers without it and explore service satisfaction for fiber users.
- Analyze streaming service satisfaction and possible tech issues causing churn.
- Reconsider pricing strategies or offer billing support to high-charge customers on electronic check.
- Improve billing service or identify the problem users facing by their feedback.

Conclusion:

This churn analysis revealed key behavioral patterns and predictors that help telecom companies anticipate customer attrition. By focusing on tenure, contract type, and service bundling, retention can be significantly improved. A machine learning-driven approach provides a scalable solution to proactively retain customers, reducing costs and boosting satisfaction.

References:

- Kaggle. (n.d.). *Telco Customer Churn Dataset*. Retrieved from https://www.kaggle.com/datasets/blastchar/telco-customer-churn

- imbalanced-learn. (n.d.). *SMOTE Oversampling – imbalanced-learn Documentation*. Retrieved from https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html

- Scikit-learn. (n.d.). *Machine Learning in Python*. Retrieved from https://scikit-learn.org/stable/

- Harvard Business Review. (2010). *Stop Trying to Delight Your Customers*. Retrieved from https://hbr.org/2010/07/stop-trying-to-delight-your-customers

Github:

https://github.com/ak-2323/Telecom-Churn-Analysis

Dashboard

https://public.tableau.com/app/profile/akshay.koladiya/viz/Telecom-Churn-Analysis_17449823682630/CustomerServiceBillingDashboard