**PROJECT REPORT**



**Customer Churn Prediction for the Telecom Industry**

**Name** : P.I.D. De Silva

**Index No** : AS2021914

**Abstract**

The telecom sector faces high customer turnover, directly impacting profitability. This project aims to develop a machine learning model that accurately predicts customer churn using historical data. By applying exploratory data analysis (EDA), feature engineering, and model comparison techniques, the project identifies key factors contributing to churn and builds a reliable predictive model to help telecom companies improve customer retention strategies. To make the insights more accessible and actionable, an interactive dashboard was also created, allowing users to explore churn patterns, feature impacts, and model performance in a visual and user-friendly format.

**Introduction**

In the telecom industry, keeping existing customers is just as important as gaining new ones. Many companies lose customers every year, which leads to lower profits and higher costs to find new users. This issue is known as **customer churn**—when users stop using a service and switch to another provider.

Today’s telecom market is very competitive. Customers have many choices and can easily change their service provider if they are unhappy with prices, network quality, or customer service. This makes it harder for telecom companies to keep customers for a long time.

Predicting which customers are likely to leave can help companies take action before it’s too late. They can offer special deals, improve services, or reach out to unhappy users. This saves money and helps grow their business.

In this project, a machine learning model was built to predict customer churn using real telecom data. By analyzing customer behavior and usage patterns, the goal is to help telecom companies understand why customers leave and how they can reduce churn in the future.

**Problem Statement**

Telecom companies face a major challenge in retaining their customers, as frequent customer churn leads to increased operational costs and reduced profits. Understanding and predicting churn is essential for building effective customer retention strategies. Churn can occur in various forms, and recognizing these scenarios helps companies respond with more targeted actions. Common types of churn include:

1. **Tariff Plan Churn** – Customers downgrade their plans (e.g., from Rs.1000 to Rs.500/month), which reduces revenue without actually leaving the service.
2. **Service Churn** – Customers stop using certain services (e.g., canceling a weekly or monthly subscription).
3. **Product Churn** – Customers switch between offerings (e.g., moving from a postpaid to a prepaid connection).
4. **Usage Churn** – Customers become inactive or stop using the service entirely, even if their account remains open.

By predicting churn in its various forms, telecom companies can proactively intervene and retain valuable users through customized offers, improved service quality, or engagement strategies.

**Objectives**

* To perform **exploratory data analysis (EDA)** to understand the key patterns and trends that influence customer churn.
* To preprocess the dataset and apply suitable techniques such as encoding, scaling, and handling class imbalance to prepare data for modeling.
* To **select and evaluate multiple machine learning models**, and identify the best-performing model based on accuracy, recall, and F1-score.
* To **train the final model** using the selected features and optimal parameters to predict churn effectively.
* To **develop an interactive dashboard** that allows users to input new customer data and receive real-time churn predictions, as well as explore model insights visually.



**Methodology**

#### 1. ****Data Collection****

The dataset used for this project was collected from a telecom service provider and includes a wide range of customer-related information. It contains both **demographic data** (such as gender, age group, and senior citizen status) and **service usage details** (such as tenure, contract type, monthly charges, internet service, and additional subscriptions like online security, tech support, and streaming services). It also includes the **Churn** column, which indicates whether a customer has left the service.

The data reflects real-world customer behavior and service interaction patterns, making it ideal for training a predictive churn model. This diverse and structured dataset enables meaningful analysis of how different factors contribute to churn. Special attention was given to ensuring the dataset was complete, relevant, and sufficiently representative of different customer segments and churn scenarios.

### ****2.Exploratory Data Analysis (EDA)****

The project began with an in-depth exploratory data analysis to understand customer behavior and detect trends linked to churn. Various visualizations such as bar plots, histograms, and heatmaps were used to analyze the distribution of key features and their correlation with the target variable (Churn).  
Significant patterns were discovered — for instance:

* Customers with **month-to-month contracts** showed a much higher likelihood of churning.
* **Shorter tenure** was closely associated with higher churn rates.
* **Higher monthly charges** also increased the probability of churn, especially when paired with limited additional services.

These findings helped shape the feature selection process and provided a clear direction for model training.

### ****3.Data Preprocessing****

To ensure the data was ready for machine learning, several preprocessing steps were applied:

* **Encoding Categorical Variables**: Categorical features were converted into numerical format using one-hot encoding and label encoding as appropriate.
* **Handling Missing Values**: Blank or non-convertible entries in the TotalCharges column were identified and converted to NaN, then handled through imputation or removal.
* **Scaling Features**: Numerical features were normalized to bring them onto a similar scale, which helps certain models perform better.
* **Balancing Classes**: The dataset was imbalanced (more non-churned than churned customers), so **SMOTEENN** was used to balance it by combining over-sampling (SMOTE) with cleaning (ENN).

### ****4.Model Selection and Evaluation****

Multiple machine learning algorithms were tested and compared using key performance metrics:

#### Models Evaluated:

* Decision Tree
* **Random Forest** (selected as the final model)

#### Evaluation Metrics:

* Accuracy
* Precision, Recall, and F1-Score
* ROC-AUC Curve
* Confusion Matrix

After rigorous evaluation, the **Random Forest Classifier** was selected as the final model due to its strong balance between precision and recall and its ability to handle complex feature interactions. It achieved an impressive **accuracy of 94.27%** on the test dataset.

### ****5.Dashboard Development****

To make the model insights and predictions more interactive and user-friendly, a **dashboard** was developed. The dashboard allows users to:

* **Input new customer data** and instantly get a churn prediction.
* **Visualize churn trends** by feature (e.g., by contract type, tenure, charges).
* **Explore feature importance** as determined by the Random Forest model.

This dashboard bridges the gap between technical analysis and business decision-making, giving telecom stakeholders a clear, visual way to understand and act on churn predictions.

#### ****Tools & Technologies****

* **Language**: Python
* **IDE**: Jupyter Notebook
* **Libraries**: pandas, numpy, matplotlib, seaborn, scikit-learn, imbalanced-learn

### ****Results****

The final model selected for this project was the **Random Forest Classifier**, which demonstrated strong predictive power with an **accuracy of 92.78%** and a high **F1-score**, indicating a balanced performance between precision and recall — especially for identifying customers who are likely to churn.

To address class imbalance in the dataset, the **SMOTEENN technique** was applied, which effectively improved model generalization and helped capture minority class (churn) behavior more accurately.

#### Key Findings from the Data:

The following features were found to have the most significant impact on customer churn:

* **Tenure**: Customers with shorter service durations were more likely to churn, indicating that newer users tend to leave if not satisfied early.
* **Contract Type**: Customers on **month-to-month contracts** showed a much higher churn rate, likely due to the flexibility to leave without penalties.
* **Monthly Charges**: Higher monthly fees were correlated with increased churn, especially when customers did not receive additional service value.
* **Electronic Check Payment Method**: Customers paying via **electronic check** had the highest churn rate, possibly due to less commitment or auto-renewal.
* **Online Security & Tech Support**: Customers who **did not subscribe to online security or tech support services** were significantly more likely to churn, suggesting these services add perceived value and improve retention.
* **Senior Citizen Status**: Interestingly, **non-senior citizens** had a higher churn rate, indicating that younger or more mobile customers may be more likely to switch providers.

These results not only validate the strength of the model but also provide actionable insights for telecom companies. By targeting high-risk customer groups — such as those without added services, those on flexible payment plans, or early-stage users — companies can design tailored retention strategies to reduce churn and enhance customer satisfaction.

#### ****Business Insights & Recommendations****

* Customers on **month-to-month contracts** with **high monthly charges** are more likely to churn.
* Offering **longer-term discounts** or **value-added services** could improve retention.
* Personalized communication should target high-risk customer profiles based on model outputs.

#### ****Limitations****

* Sentiment analysis from call center logs or support chats was not included due to data unavailability.
* Real-time churn prediction was not implemented.
* External market factors (e.g., competitor offers) were not modeled.

#### ****Conclusion****

This project successfully built a churn prediction model using telecom customer data. By identifying at-risk users, telecom providers can take proactive steps to reduce churn. The findings and recommendations can be directly applied in customer retention strategies. Future work may integrate NLP techniques to include customer feedback analysis for even more robust predictions.

**References**

* <https://youtu.be/bDhvCp3_lYw?si=Mnhm0Tn-QHXiw_BW>
* <https://youtu.be/38SUUaMX5Rg?si=1miNYvx-ftl8mTqQ>