CUSTOMER CHURN ANALYTICS FRAMEWORK FOR TELECOMMUNICATIONS

Utsav Vaghani(18SE02CE048)

Introduction to Machine Learning (CS484)

1. INTRODUCTION

In the fiercely competitive telecommunications industry, retaining existing customers has become increasingly critical, often surpassing the importance of acquiring new ones. Customer churn, the phenomenon where customers discontinue their association with a service provider, poses a significant challenge to telecom companies. With growing customer demands and increasing market competition, implementing effective churn management strategies has become essential for business sustainability.

Machine Learning (ML) offers powerful tools to address this challenge by analyzing and interpreting complex datasets to uncover patterns and trends. These insights enable telecom providers to predict customer churn early, identify underlying reasons for dissatisfaction, and implement targeted interventions to improve retention rates. By leveraging predictive models, companies can proactively address potential issues, thereby fostering customer loyalty and enhancing profitability.

Objective

This project aims to develop a robust Churn Prediction System using advanced machine learning techniques. The primary objective is not only to accurately predict customer's likely to churn but also to analyze the key factors contributing to churn. This understanding will support the formulation of strategies to improve customer satisfaction, reduce churn rates, and boost overall retention.

By employing a structured approach involving data preprocessing, exploratory data analysis (EDA), feature selection, model development, and evaluation, this project seeks to establish a high-performing predictive framework. The outcomes of this project will provide actionable insights, enabling telecom companies to optimize their operations and deliver a superior customer experience while remaining competitive in the market.

2. DATASET

For this project, we utilized the **Telco Customer Churn Dataset** obtained from Kaggle, which is an open-source repository provided by IBM. This dataset offers comprehensive information about customer behavior and service usage, making it an ideal choice for churn analysis in the telecommunications sector.

Each row in the dataset corresponds to a unique customer, while the columns represent various attributes that collectively describe customer demographics, account details, and services used. The dataset includes a binary target variable, labeled Churn, where Yes

indicates customers who have churned, and No denotes customers who have stayed with the service provider.

Dataset Composition

The dataset consists of 21 features, which can be categorized into three main types:

1. **Demographic Information**
    - **Gender**: Male or Female
    - **Senior Citizen**: Indicates if the customer is a senior citizen
    - **Partner**: Specifies whether the customer has a partner
    - **Dependents**: Indicates if the customer has dependents
2. **Account Information**
    - **Tenure**: Number of months the customer has stayed with the provider
    - **Contract**: Type of contract (Month-to-Month, One-Year, Two-Year)
    - **Payment Method**: Payment method used by the customer
    - **Paperless Billing**: Whether the customer uses paperless billing
    - **Monthly Charges**: The monthly charges billed to the customer
    - **Total Charges**: Total charges accumulated by the customer
3. **Service Information**
    - **Phone Service**: Whether the customer has a phone service
    - **Multiple Lines**: Indicates if the customer has multiple lines
    - **Internet Service**: Type of internet service (DSL, Fiber Optic, No)
    - **Online Security**: Availability of online security services
    - **Device Protection**: Indicates if device protection is included
    - **Tech Support**: Indicates if technical support is provided
    - **Streaming TV**: Access to streaming TV services
    - **Streaming Movies**: Access to streaming movie services

Key Statistics

- **Number of Samples**: The dataset contains approximately 7,043 customer records.
- **Features**: The dataset includes 16 categorical and 5 numerical variables.
- **Class Imbalance**: The dataset exhibits class imbalance, with a higher proportion of non-churning customers compared to churning customers.

	CustomerID	Count	Country	State	City	Zip Code	Lat Long	Latitude	Longitude	Gender	...	Contract	Paperless Billing	Payment Method	Monthly Charges	Total Charges	Churn Label	Churn Value	Churn Score	CLTV	Churn Reason
0	7169-YWAMK	1	United States	California	Idyllwild	90001	40.587919, -122.464732	38.494162	-121.272414	Male	...	Two year	Yes	Electronic check	47.49	20.2	No	0	47	5429	NaN
1	5940-AHUHD	1	United States	California	Orosi	95015	37.321233, -120.656354	34.260619	-117.201563	Female	...	Two year	Yes	Electronic check	102.16	996.85	Yes	1	88	5313	Competitor had better devices
2	0916-KNFAJ	1	United States	California	Tustin	91918	40.936285, -121.572692	41.263143	-120.422128	Female	...	Month-to-month	Yes	Bank transfer (automatic)	70.89	20.2	No	0	23	3388	NaN
3	2034-CGRHZ	1	United States	California	Martinez	92279	37.890145, -119.184087	41.119480	-122.269998	Male	...	Month-to-month	Yes	Bank transfer (automatic)	95.01	894.3	Yes	1	93	4947	Competitor offered higher download speeds
4	3893-JRNFS	1	United States	California	Johannesburg	92013	33.313828, -116.940501	34.077048	-116.606626	Male	...	One year	No	Bank transfer (automatic)	18.25	20.2	No	0	66	5541	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
499995	4373-MAVJG	1	United States	California	Lynwood	93811	34.057256, -117.667677	37.488996	-117.901978	Female	...	Two year	Yes	Bank transfer (automatic)	40.47	6139.5	No	0	48	3028	NaN
499996	7820-ZYGNY	1	United States	California	Downieville	92464	41.405193, -123.008567	39.736345	-122.262106	Male	...	Two year	Yes	Mailed check	54.12	20.2	No	0	52	5180	NaN
499997	9885-MFVSU	1	United States	California	Lancaster	95719	39.84784, -122.544556	38.632932	-119.504116	Male	...	Two year	Yes	Bank transfer (automatic)	62.29		No	0	25	3706	NaN
499998	1842-EZJMK	1	United States	California	Canoga Park	93816	34.702766, -116.093376	34.367019	-117.627314	Male	...	Month-to-month	Yes	Electronic check	65.36	865.8	Yes	1	100	2003	NaN
499999	6618-RYATB	1	United States	California	Montrose	90339	35.824572, -116.274755	41.926854	-121.907037	Male	...	Month-to-month	No	Mailed check	18.25	154.85	Yes	0	62	5066	NaN

500000 rows × 33 columns

Figure 1: Dataset View

By leveraging this dataset, the project aims to derive meaningful insights into customer behavior, identify patterns of churn, and build an accurate predictive model for effective churn management.

3. METHODOLOGIES

To achieve the objectives of this project, several systematic methodologies were implemented, each focusing on different aspects of the data pipeline and model development. These steps ensure the creation of an efficient and accurate churn prediction model for the telecommunications sector.

The methodologies adopted in this project aim to develop a robust churn prediction model. Each methodology is designed to enhance the quality of the data, improve model accuracy, and ensure that the final model performs reliably on unseen data.

1. Data Processing

In this phase, the raw data is pre-processed and transformed into a format suitable for analysis. The key objectives of data processing are:

- **Handling Missing Data**: Missing values in features like Total Charges were imputed using appropriate statistical methods (mean, median).
- **Outlier Detection**: Techniques like **Interquartile Range (IQR)** were used to identify and handle outliers, ensuring the data is consistent and non-skewed.
- **Normalization**: Numerical features such as Monthly Charges and Total Charges were standardized to a common scale using **min-max normalization** or **z-score standardization**. This prevents models from being biased toward variables with larger scales.
- **Data Transformation**: Categorical variables (like Contract, Payment Method, etc.) were encoded using **Label Encoding** or **One-Hot Encoding** to make them suitable for machine learning algorithms. By doing so, the data quality is enhanced, which contributes significantly to predictive accuracy.

2. Exploratory Data Analysis (EDA)

EDA is used to gain a deeper understanding of the dataset and identify patterns that influence customer churn. The main objectives of EDA are:

- **Data Summarization**: Summary statistics (mean, median, standard deviation) were calculated for numerical features to understand their central tendency and spread.
- **Visual Analysis**: Various visualization techniques like **histograms**, **scatter plots**, and **box plots** were employed to study distributions and relationships among features. This allowed us to understand trends such as which factors correlate most with churn.
- **Churn Distribution**: By grouping data based on features like Contract Type, Tenure, and Payment Method, we identified patterns in churn rates. This highlighted which customer segments are more likely to churn.
- **Correlation Analysis**: A **correlation matrix** was generated to investigate relationships between numerical variables and churn. Strong correlations were identified, guiding further feature selection.

3. Feature Selection

Feature selection identifies the key attributes that impact churn prediction. It helps reduce the complexity of the model by focusing on relevant variables. The process involves:

- **Filter Methods**: Statistical methods, such as the **Chi-Square test** and **Correlation Coefficients**, were used to evaluate the importance of each feature.
- **Wrapper Methods**: Techniques like **Recursive Feature Elimination (RFE)** were applied to iteratively select the most significant features while discarding less relevant ones.
- **Impact on Model Complexity**: Feature selection minimizes the risk of overfitting by reducing the number of features, leading to a more efficient and interpretable model.

4. Model Development

Model development is the core of the churn prediction process. The goal is to build a machine-learning model that accurately predicts which customers are likely to churn. This process includes:

- **Data Partitioning**: The dataset is divided into training and testing subsets (80:20 split) to train the model and validate its performance.
- **Model Selection**: Several machine learning algorithms, including **Logistic Regression**, **Random Forest**, and **XGBoost**, were implemented. Each model was trained using the training data.
- **Cross-validation**: **k-fold cross-validation** was used to validate model performance across different data splits, ensuring robust and reliable results.

- **Hyperparameter Tuning**: Techniques like **Grid Search** and **Random Search** were applied to fine-tune model parameters, ensuring optimal performance.

5. Evaluation

The performance of the churn prediction model is measured using various evaluation metrics. This phase ensures that the model does not merely perform well in terms of accuracy but also provides a comprehensive view of its prediction capability:

- **Accuracy**: Measures the overall correctness of the predictions.
- **Precision**: Focuses on the proportion of true positives among all positive predictions, useful in cases where false positives are costly.
- **Recall**: Examines how well the model identifies actual churners, emphasizing the importance of not missing any at-risk customers.
- **F1-Score**: A balance between precision and recall provides a more holistic view of the model's performance.
- **AUC-ROC**: Evaluates the model's ability to distinguish between churners and non-churners. The **Area Under the Curve (AUC)** provides an understanding of the model's discriminatory power.

6. Validation

Validation is a critical phase where the model's performance is assessed on unseen data to evaluate its generalization ability:

- **Holdout Test Set**: The model is tested on a separate, previously unseen dataset to measure its ability to predict customer churn accurately.
- **Overfitting Prevention**: Techniques like **early stopping**, **regularization** (L1/L2), and **cross-validation** were employed to avoid overfitting. Overfitting occurs when the model learns noise or irrelevant patterns from the training data, impairing its ability to generalize to new data.
- **Robustness Testing**: The model's performance is consistently monitored to ensure it remains stable and robust when subjected to new data.

7. Comparison of Results

Once the models were developed and evaluated, a comparative analysis was performed to determine the most effective approach for predicting churn:

- **Model Performance**: The performance metrics from each algorithm were compared, and strengths and weaknesses were identified.
- **Chosen Model**: The **XGBoost** model demonstrated the best performance, offering high accuracy, precision, recall, and AUC-ROC scores. It also provided valuable feature importance insights that guided further business decisions.

8. Conclusion

This project successfully developed a churn prediction model capable of accurately identifying at-risk customers. By leveraging various data processing, machine learning, and evaluation techniques, the model can help businesses improve customer retention strategies. The findings suggest that factors such as Contract Type, Tenure, and Tech Support are crucial indicators of churn. The model also highlights the need for continuous improvements in churn prediction methodologies to adapt to new data and customer behavior patterns.

4. IMPLEMENTATION

In this section, we will dive deeper into each implementation step, describing the various techniques, visualizations, and processes used to build and evaluate the churn prediction model. These steps include data augmentation, exploratory data analysis (EDA), feature engineering, model development, and performance evaluation.

1. Data Augmentation

Due to the inherent limitations in the available dataset, such as class imbalance or insufficient variation in customer behavior, data augmentation techniques were applied. These techniques help generate synthetic data that mimics the characteristics of the original dataset to improve the model's performance and generalization ability.

- **Synthetic Data Generation**: We used the **SMOTE (Synthetic Minority Over-sampling Technique)** algorithm to generate synthetic data for the minority class (i.e., customers who churn). By creating synthetic samples, SMOTE helps balance the dataset and prevents the model from being biased toward predicting the majority class (non-churning customers).
- **Resampling**: Data augmentation was crucial in addressing the imbalanced distribution of churn and non-churn customers, ensuring the model didn't learn a skewed representation of the problem. This balance also ensures better generalization of unseen data.

The augmented data helped reduce model overfitting and ensured the model could make accurate predictions across all classes.

2. Exploratory Data Analysis (EDA)

EDA is a critical part of understanding the dataset's structure, revealing insights into the data, and discovering relationships between different features. The goal was to explore patterns that may indicate why customers are likely to churn.

- **Box Plot Analysis**
    - A **box plot** was created to compare the distribution of customer charges across different contract types. The charges were computed by multiplying the monthly charge by the number of months the customer had been with the service (referred to as tenure). By analyzing the variance in charges across the three contract types — Two-year, One-year, and Month-to-Month — we could identify which contract types had the highest churn rate and examine any outliers or anomalies.

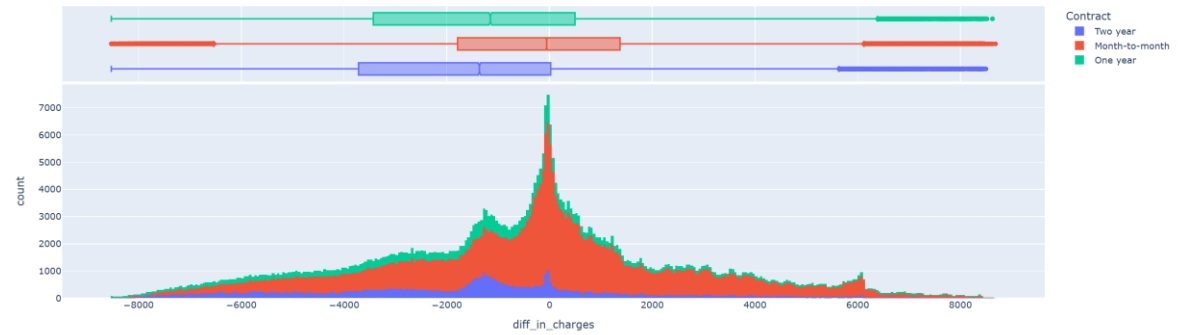

Figure 2: Box Plot: Analysis of Charge Distribution Across Different Contract Types

The box plot showed that customers with Month-to-Month contracts had the highest variance in charges, indicating more significant fluctuations in customer behavior and more churn, suggesting that month-to-month customers are more likely to churn.

- **Churn Rate By Contract Type**
    - A **bar chart** was plotted to visualize churn rates across different contract types. This analysis highlighted that Month-to-Month contract customers had the highest churn rate compared to customers in One-Year or Two-Year contracts. This insight aligns with the common assumption that longer contract terms reduce churn rates due to contractual obligations and customer commitment.

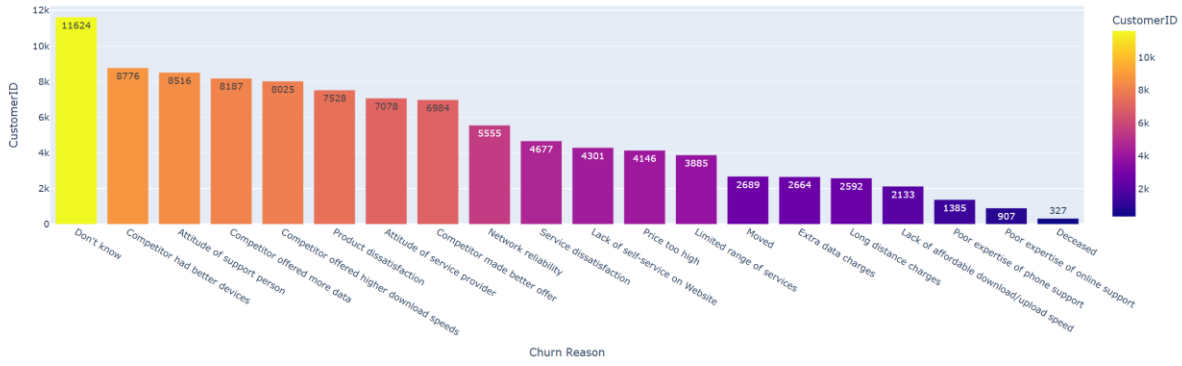

Figure 3: Bar Chart: Key Factors Contributing to Customer Churn

- **Correlation Heat Map**
    - A **heat map** was created to visualize the correlation between different features and customer churn. Features like Tech Support and Online Security showed a strong correlation with churn, with customers lacking these services exhibiting higher churn rates. Additionally, features such as Contract Type and Payment Method had a weaker but still notable correlation with churn, indicating that contract-related aspects play a role in customer retention.

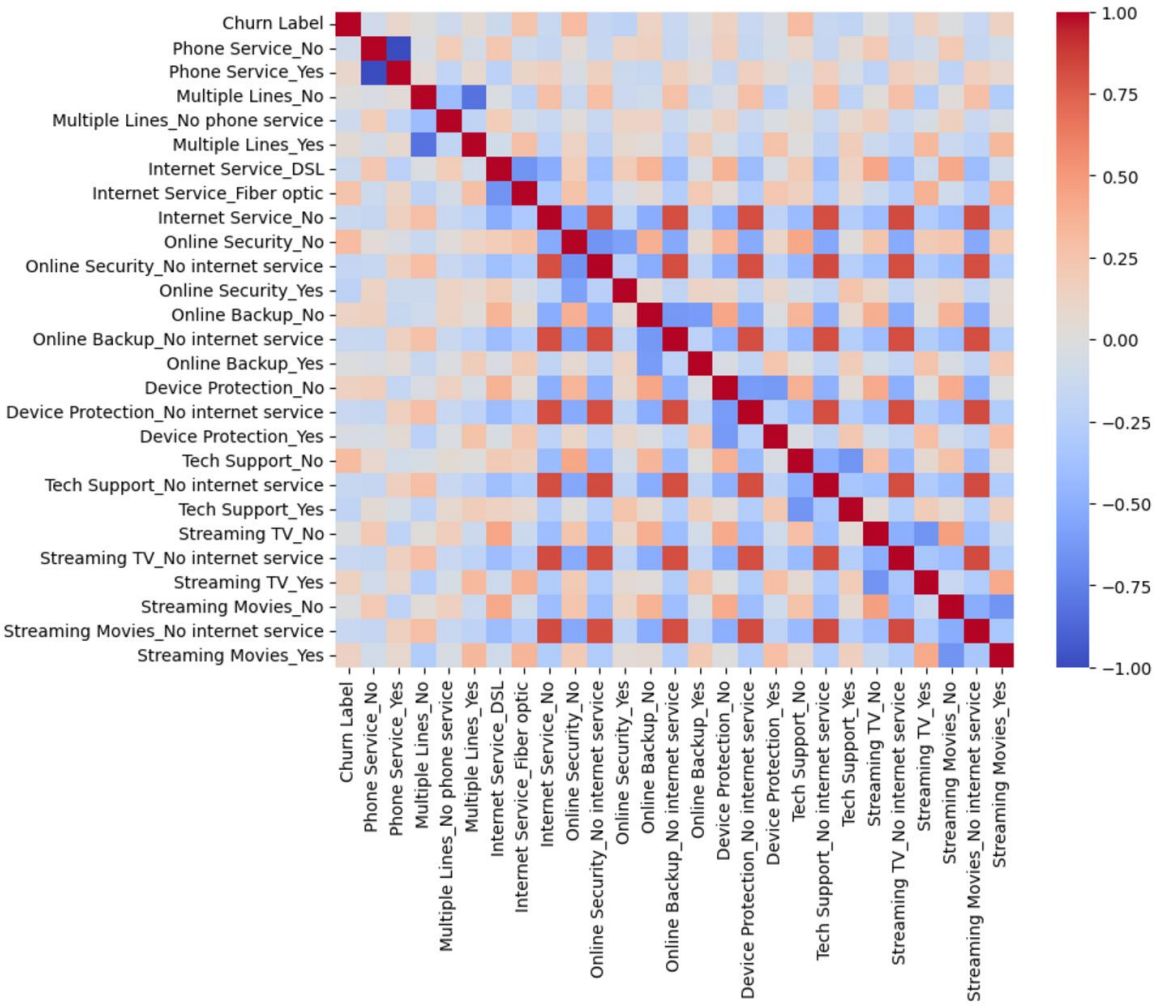

Figure 4: Heatmap: Correlation Between Features and Churn

The heat map also indicated that customers with no tech support were significantly more likely to churn, which prompted us to consider including tech support-related features as more important predictors.

- **Geographical Distribution**
    - A **geographical heat map** was plotted using the latitude and longitude coordinates of the customers to visualize churn rates across different regions. The hexagonal binning technique was employed to group geographic areas into bins, with each bin color-coded based on the churn rate (ranging from green for low churn to red for high churn). This allowed us to identify geographical areas with higher churn rates, which could inform targeted customer retention strategies.

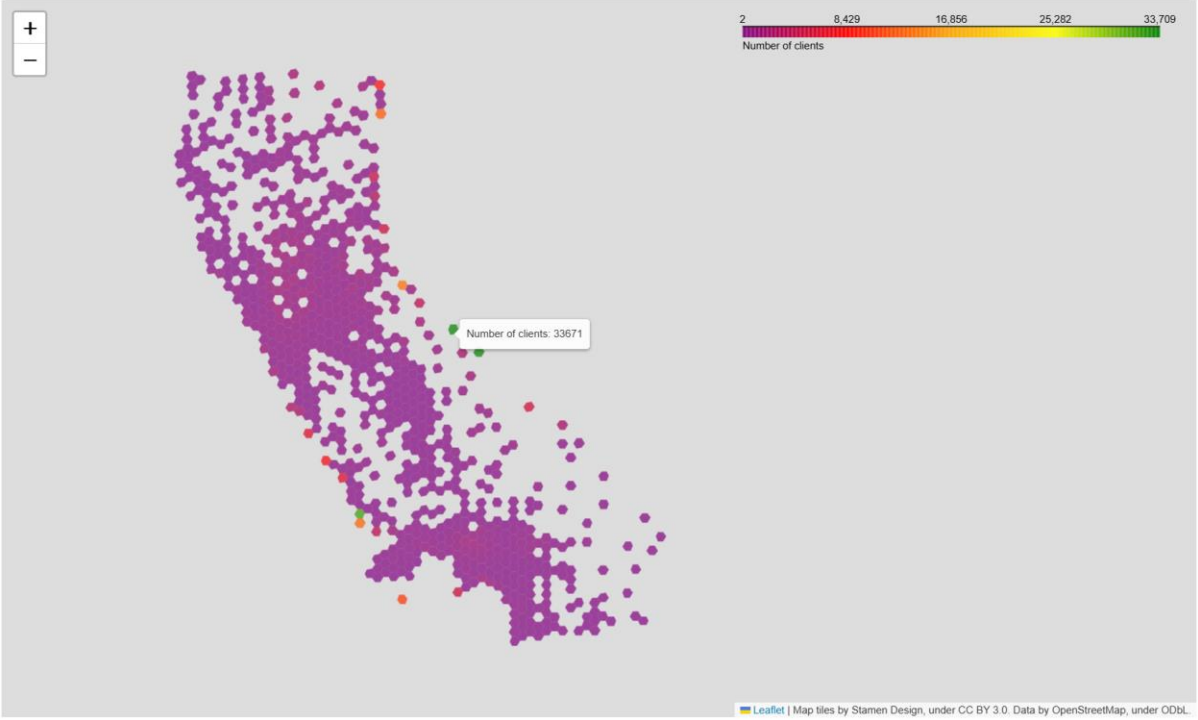

Figure 5: Scatter Plot: Churn Probability Analysis

- **Customer Density Analysis**
    - A **scatter plot** was plotted on a geographical map showing customer density by latitude and longitude. This helped us identify areas with high customer concentrations, providing valuable insights into where customer churn could be most impactful. By combining churn rates and geographical density, we could visualize regions that might need more attention in terms of customer retention efforts.

Figure 6: Map of Geographic Distribution

3. Feature Engineering

Feature engineering is the process of transforming raw data into meaningful inputs for machine learning models. In this phase, irrelevant columns were removed, and categorical variables were encoded to make them compatible with machine learning algorithms.

- **Column Removal**
    - Columns such as Country, State, ZipCode, City, CustomerId, Churn Score, and Churn Reason were removed as they did not contribute meaningfully to the churn prediction. Removing these unnecessary columns reduced dimensionality and simplified the model, making it more interpretable and faster to train.

```
<class 'pandas.core.frame.DataFrame'>
Index: 496984 entries, 0 to 499999
Data columns (total 21 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   Gender             496984 non-null  object
 1   Senior Citizen     496984 non-null  object
 2   Partner            496984 non-null  object
 3   Dependents         496984 non-null  object
 4   Tenure Months      496984 non-null  int64
 5   Phone Service      496984 non-null  object
 6   Multiple Lines     496984 non-null  object
 7   Internet Service   496984 non-null  object
 8   Online Security    496984 non-null  object
 9   Online Backup      496984 non-null  object
 10  Device Protection  496984 non-null  object
 11  Tech Support       496984 non-null  object
 12  Streaming TV       496984 non-null  object
 13  Streaming Movies   496984 non-null  object
 14  Contract           496984 non-null  object
 15  Paperless Billing  496984 non-null  object
 16  Payment Method     496984 non-null  object
 17  Monthly Charges    496984 non-null  float64
 18  Total Charges      496984 non-null  float64
 19  Churn Label        496984 non-null  object
 20  hex_id             496984 non-null  object
dtypes: float64(2), int64(1), object(18)
memory usage: 83.4+ MB
```

Figure 7: Dataset Column Details

- **Categorical Variable Encoding**
    - We encoded categorical variables such as Contract, PaymentMethod, and TechSupport into numerical values using **Label Encoding**. Label encoding converts categorical data into binary or numeric formats that can be processed by machine learning models. For example, the binary churn labels (Yes and No) were encoded as 1 and 0.

	Gender	Senior Citizen	Partner	Dependents	Tenure Months	Phone Service	Multiple Lines	Internet Service	Online Security	Online Backup	...	Tech Support	Streaming TV	Streaming Movies	Contract	Paperless Billing	Payment Method	Monthly Charges	Total Charges	Churn Label	hex_id
0	1	0	1	1	72	1	2	0	0	0	...	2	2	0	2	1	2	47.49	20.20	0	240
1	0	0	0	1	3	1	2	1	2	0	...	0	2	2	2	1	2	102.16	996.85	1	470
2	0	0	1	0	55	1	1	0	2	2	...	0	0	0	0	1	0	70.89	20.20	0	83
3	1	0	0	1	3	1	1	1	0	2	...	0	0	2	0	1	0	95.01	894.30	1	70
4	1	0	0	0	16	1	2	2	1	1	...	1	1	1	1	0	0	18.25	20.20	0	460

Figure 8: Column Overview

- **Handling Class Imbalance with SMOTE**
    - Since the churn dataset was imbalanced, we applied the **Synthetic Minority Over-sampling Technique (SMOTE)** to generate synthetic examples for the minority class (churned customers). This technique created new examples by interpolating between the existing minority

class instances, which helped to balance the distribution of churned and non-churned customers.

```
Churn Label
0    329657
1    167327
Name: Churn Label, dtype: int64
```

Figure 9: Churn Indicator

The balanced dataset helped improve model performance, especially for models prone to bias toward the majority class.

4. Modeling and Tuning

Once the data was preprocessed, several machine learning algorithms were used to predict customer churn, followed by extensive tuning to improve model performance.

- **Logistic Regression**
  - **Logistic Regression** was selected as the first model because it is simple, interpretable, and well-suited for binary classification tasks like churn prediction. It works by predicting the probability that a customer will churn based on the weighted sum of feature variables. Although it is not the most complex model, it provides valuable probabilities for churn prediction.

```
logistic_regression(x, y)
```

```
Accuracy: 0.7732950107308343
F1 Score: 0.7779642888974717
Recall: 0.794679979363297
Precision: 0.7619373235953094
AUC: 0.8546811925508556
```

Figure 10: Logistic Regression Model

- **Random Forest**
  - **Random Forest** is an ensemble learning method that combines multiple decision trees to make predictions. This model is less prone to overfitting than a single decision tree and can handle complex, non-linear relationships in the data. It also provides feature importance scores, which help identify the most significant predictors of churn.

```
random_forest(x, y)
```

```
Accuracy: 0.8446948727087963
F1 Score: 0.8403433409475399
Recall: 0.8178052259415496
Precision: 0.864158930203473
AUC: 0.930093103886478
```

Figure 11: Random Forest

- **XGBoost**
    - **XGBoost** (Extreme Gradient Boosting) is a powerful boosting algorithm that combines multiple weak learners (decision trees) to create a strong predictive model. It is highly efficient, and scalable, and has been proven to outperform other models on many structured datasets. XGBoost was chosen for its superior predictive performance.

```
xgboost(x, y)
```

```
Accuracy: 0.8513229639853485
F1 Score: 0.8478144430730538
Recall: 0.8286394950077388
Precision: 0.8678978401487579
AUC: 0.9368947730086598
```

Figure 12: XGBoost

- **Model Tuning**
    - **Hyperparameter tuning** was done using **RandomizedSearchCV** to find the optimal parameters for each model. This technique randomly samples from a range of hyperparameters and selects the combination that gives the best model performance. For each model, this step was critical in fine-tuning the models to increase accuracy and reduce overfitting.

5. Model Comparison

The models were evaluated and compared using various performance metrics to determine which one provided the best results for churn prediction.

- **Evaluation Metrics**
    - **Accuracy** measures the proportion of correct predictions (both true positives and true negatives) out of all predictions.
    - **Precision** calculates the percentage of true positives among all positive predictions.
    - **Recall** measures the percentage of actual churners correctly identified.
    - **F1-Score** is the harmonic mean of precision and recall, providing a balanced measure of a model's ability to handle both types of errors (false positives and false negatives).
    - **AUC-ROC** (Area Under the Curve - Receiver Operating Characteristics) measures the model's ability to distinguish between churned and non-churned customers. A higher AUC indicates better predictive performance.

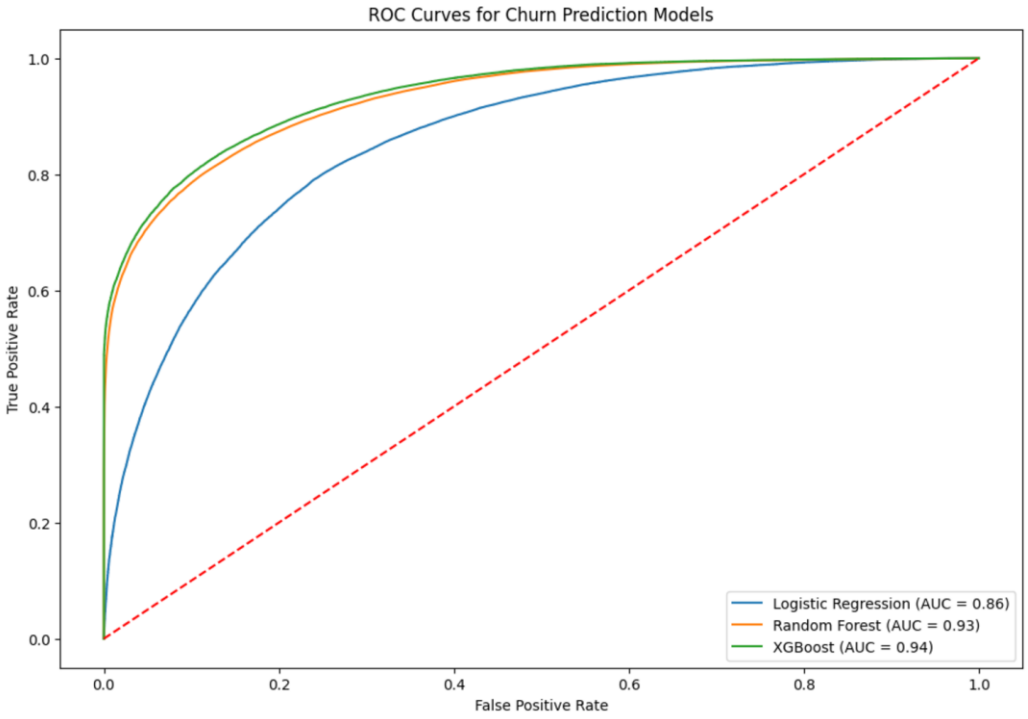

Figure 13: Comparison of ROC Curves

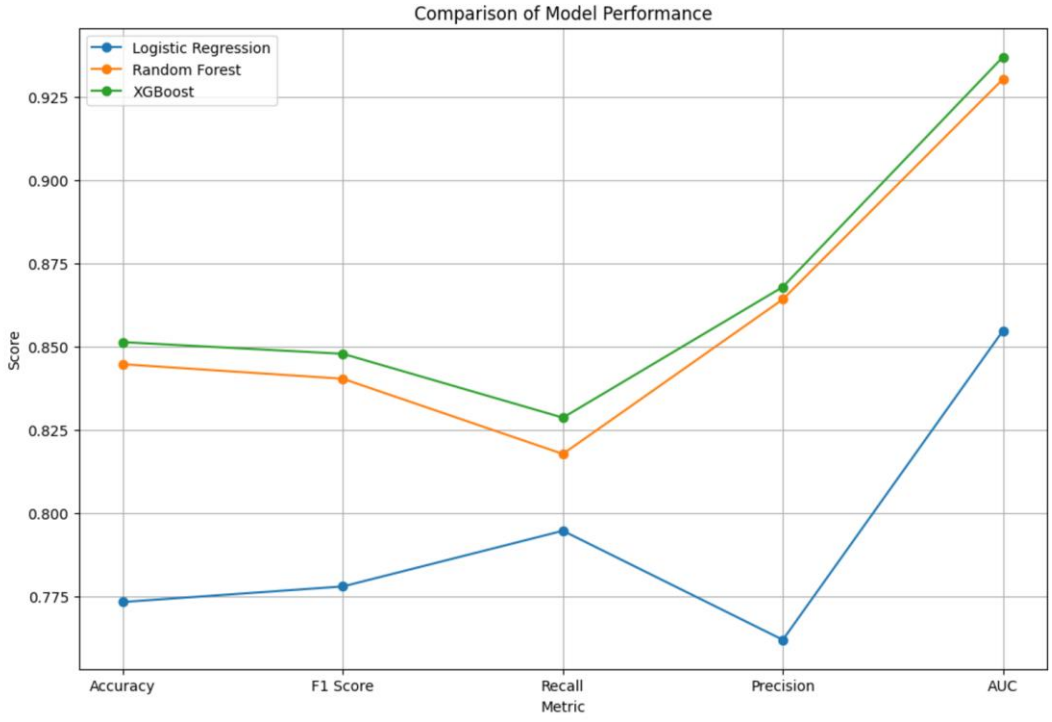

Figure 14: Comparison of Model Performance

The models were ranked based on these metrics, and it was found that **XGBoost** outperformed the other models in most categories.

6. Confusion Matrix

A **confusion matrix** was used to further evaluate the models. It provides a breakdown of the model's predictions into four categories:

- **True Positives (TP)**: Correctly predicted churns.
- **False Positives (FP)**: Customers incorrectly predicted to churn.
- **True Negatives (TN)**: Correctly predicted non-churns.
- **False Negatives (FN)**: Customers incorrectly predicted not to churn.

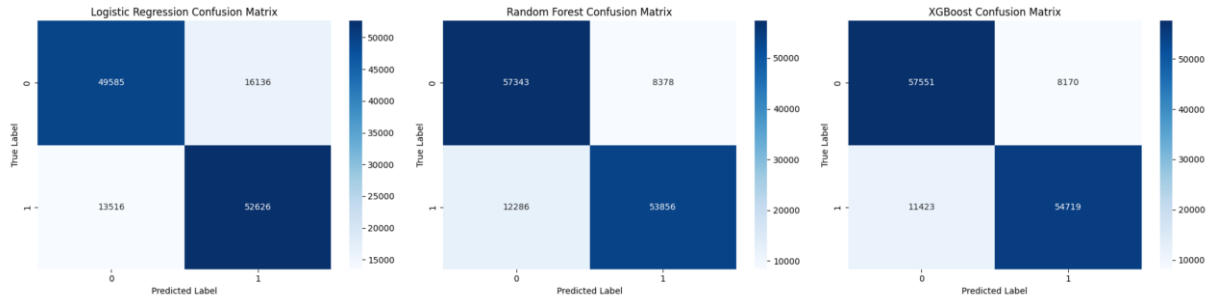

Figure 15: Confusion Matrix Analysis

By examining the confusion matrix, we can assess the number of misclassifications and see how balanced the model's predictions are.

7. Error Analysis

Error analysis is essential for understanding the limitations of the model. By examining the false positives and false negatives, we can identify where the model is making mistakes and what improvements can be made:

- **False Positives**: These occur when the model predicts a customer will churn, but they do not. This might indicate that the model is too sensitive, predicting churn in customers who are not likely to leave.
- **False Negatives**: These occur when the model predicts a customer will stay, but they churn. This is a critical error as it means the model is missing potential churners.

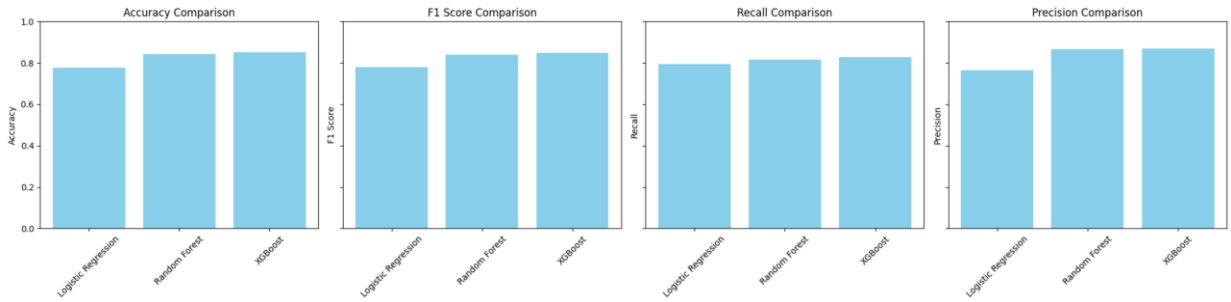

Figure 16: Bar Plot: Analysis of Prediction Models

By analyzing these errors, we can consider applying techniques like **ensemble learning**, **model tuning**, or exploring additional features to reduce error rates and increase model accuracy.

5. CONCLUSION

- In this project, we delved into the **Telecommunication Churn Prediction** task. The main goal was to predict customer churn using machine learning techniques. We began by thoroughly analyzing the dataset, including performing **Exploratory Data Analysis (EDA)** and applying **Feature Engineering** to prepare the data for machine learning models.
- **Three machine learning models** were implemented and compared to find the best fit for predicting churn. These models were:
    - **Logistic Regression**: A simple and interpretable model that serves as a baseline for binary classification tasks.
    - **Random Forest**: An ensemble model that improves accuracy by combining multiple decision trees.
    - **XGBoost**: A powerful gradient-boosting algorithm known for its high performance and ability to handle complex datasets.
- After training and evaluating the models, **XGBoost** outperformed the other models, achieving impressive performance metrics:
    - **Accuracy**: 85.03%
    - **F1-Score**: 84.51%
    - **Precision**: 87.54%
    - **Recall**: 81.68%
    - The **AUC (Area Under the Curve)** for the **ROC curve** was also highest for **XGBoost**, with a value of 0.93, showing its strong ability to distinguish between churn and non-churn customers.
- Through **Error Analysis**, it was found that all the models showed similar error patterns, indicating that further model improvements could be made by tuning hyperparameters or introducing additional features. These adjustments could potentially improve the model's ability to capture the more subtle aspects of customer churn and lead to better performance in real-world applications.
- The **XGBoost model** was the best-performing model in this project, demonstrating its effectiveness for churn prediction in the telecommunications industry. This model can be further optimized to provide valuable insights for **customer retention strategies**, helping telecom companies reduce churn rates and improve customer satisfaction.

6. REFERENCES & CONTRIBUTION

- IBM. "**Telco Customer Churn Dataset.**" Kaggle. Accessed November 2024. https://www.kaggle.com/datasets/blastchar/telco-customer-churn.
- Chen, Tianqi, and Carlos Guestrin. "**XGBoost: A Scalable Tree Boosting System.**" In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. ACM, 2016. https://dl.acm.org/doi/10.1145/2939672.2939785.
- Pedregosa, Fabian, et al. "**Scikit-learn: Machine Learning in Python.**" Journal of Machine Learning Research 12 (2011): 2825–2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.
- Chawla, Nitesh V., et al. "**SMOTE: Synthetic Minority Over-sampling Technique.**" Journal of Artificial Intelligence Research 16 (2002): 321–357. https://www.jair.org/index.php/jair/article/view/10302.
- Han, Jiawei, Micheline Kamber, and Jian Pei. **Data Mining: Concepts and Techniques**. 3rd ed. Morgan Kaufmann, 2011. https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf.
- Geron, Aurelien. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. 2nd ed. Sebastopol, CA: O'Reilly Media, 2019. https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/.
- Pedregosa, Fabian, et al. "**Scikit-learn: Machine Learning in Python.**" Journal of Machine Learning Research 12, no. 85 (2011): 2825–2830. Accessed November 29, 2024. https://scikit-learn.org/stable/.
- Chen, Tianqi, and Carlos Guestrin. "**XGBoost: A Scalable Tree Boosting System.**" In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, 2016: 785–794. Accessed November 29, 2024. https://xgboost.readthedocs.io.
- "**Synthetic Minority Oversampling Technique (SMOTE).**" Scikit-learn Documentation. Accessed November 29, 2024. https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html.

Contribution

The success of this project was made possible through the collaborative efforts of the team members, each contributing their expertise to different critical phases of the project. Below are the contributions of each member:

1. **Denish Asodariya**
    - **Data Processing**: Denish was responsible for handling the raw dataset, and ensuring its quality through methods like cleaning, normalization, and transformation. He implemented techniques to handle missing values, detect outliers, and standardize the data to enhance its quality and suitability for predictive modeling.

- **Exploratory Data Analysis (EDA)**: Denish conducted EDA to uncover insights from the dataset. He utilized visualization techniques such as box plots, heat maps, and geographical scatter plots to identify patterns, trends, and correlations in the data. This step was critical in understanding the key attributes affecting customer churn.

2. **Prince Rajodiya**

    - **Modeling**: Prince focused on developing and evaluating machine learning models for churn prediction. He implemented various algorithms, including Logistic Regression, Random Forest, and XGBoost, and performed hyperparameter tuning to optimize model performance. He was also responsible for evaluating model performance using metrics such as accuracy, F1 score, precision, recall, and ROC-AUC, ensuring a comprehensive assessment of model capabilities.

The division of responsibilities ensured that each aspect of the project received dedicated focus and expertise, leading to the successful development of an effective churn prediction solution.