CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION TO TELECOM CUSTOMER LOSS PREDICTION USING RANDOM FOREST

• **Telecom Customer Retention:** In the competitive telecommunications sector, retaining subscribers—often referred to as customer retention—has become as crucial as acquiring new ones. The industry is characterized by market saturation, making it imperative for companies to focus on minimizing customer churn.

• **Understanding Customer Churn:** To effectively prevent customer loss, telecom companies must first comprehend the underlying causes of churn. Common reasons include dissatisfaction with service quality, enticing offers from competitors, or inadequate customer support. Data analytics plays a pivotal role in identifying these churn patterns and predicting potential customer exits. By anticipating these issues, companies can intervene proactively with tailored offers, enhanced service quality, or improved support to retain their customers.

• **Enhancing Customer Experience:** Delivering a superior customer experience is fundamental to retention. This involves providing consistent and reliable service, maintaining responsive customer support, and offering value-added services. Implementing loyalty programs, discounts, and exclusive deals can further incentivize customers to remain with the company.

• **Technological Advancements and Network Infrastructure:** Keeping pace with technological advancements and continuously upgrading network infrastructure ensures that customers receive high-quality service. Regularly gathering customer feedback and engaging with customers helps address their needs and concerns promptly, contributing to improved satisfaction and reduced churn.

• **Effective Communication:** Transparent and frequent communication is essential. Keeping customers informed about new features, service improvements, and potential issues helps build trust and fosters a sense of transparency. By employing these strategies, telecom companies can effectively minimize churn, build customer loyalty, and sustain a profitable customer base amidst industry competition.

1.2 NEED FOR CUSTOMER LOSS PREDICTION

1. **Revenue Loss:** Customer churn results in considerable revenue loss, directly impacting profitability and financial stability.
2. **Acquisition Costs:** The expense of acquiring new customers is generally higher than the cost of retaining existing ones. Reducing churn mitigates the need for costly acquisition efforts.
3. **Competition:** High churn rates can erode market share, giving competitors an opportunity to capture dissatisfied customers.

4. **Brand Reputation:** Frequent churn can harm a company's reputation, leading to a negative perception among potential and current customers.
5. **Data Insights:** Analyzing churn data provides valuable insights into areas that require improvement, helping to refine service offerings and customer interactions.
6. **Proactive Action:** Early detection of potential churn allows for the implementation of targeted retention strategies, reducing the likelihood of customer exit.
7. **Customer Satisfaction:** Lowering churn rates contributes to overall customer satisfaction by addressing and mitigating the factors that lead to dissatisfaction.
8. **Resource Allocation:** Efficient churn detection aids in optimizing resource allocation, ensuring that efforts and investments are directed where they are most needed.

1.3 AIM OF THIS PROJECT

The primary goal of telecom customer loss (or churn) prediction is to identify customers who are at risk of discontinuing their service within a specified period. By anticipating potential churn, telecom companies can take proactive measures to retain these customers. Specifically, the objectives are:

- **Increase Customer Retention:** Identifying customers at risk of churn enables the implementation of targeted retention strategies, such as personalized offers, discounts, or enhanced customer service.
- **Reduce Revenue Loss:** Preventing customer churn helps to maintain a consistent revenue stream and minimizes the costs associated with acquiring new customers to replace those who leave.
- **Improve Customer Satisfaction:** Gaining insights into the factors contributing to churn allows telecom companies to improve their services and address customer pain points, leading to greater overall satisfaction.
- **Optimize Marketing Strategies:** By understanding which customers are likely to churn, marketing efforts can be more effectively focused on retention, rather than broadly targeting the entire customer base.
- **Enhance Competitive Advantage:** Effective churn prediction provides a competitive edge by helping companies retain a stable customer base and continuously improve their services based on predictive insights.

CHAPTER 2

LITERATURE REVIEW

2.1 HISTORY OF TELECOM CUSTOMER LOSS PREDICTION USING RANDOM FOREST

Telecom customer churn prediction has long been a critical area of research due to its substantial impact on customer retention and revenue. Various studies and projects have explored effective models for predicting customer churn. Some notable contributions include:

- **Research by Huang et al. (2015):**
    - **Objective:** To apply deep learning techniques for churn prediction.
    - **Data:** Large-scale data from a Chinese telecom operator, including call detail records (CDRs) and customer profiles.
    - **Results:** The deep learning model demonstrated superior performance compared to traditional machine learning models, highlighting the effectiveness of advanced neural networks in capturing complex patterns.
- **Study by Idris et al. (2012):**
    - **Objective:** To develop a hybrid model combining decision trees and support vector machines (SVM) for churn prediction.
    - **Data:** Data from a telecom company, including customer demographics, billing information, and service usage patterns.
    - **Results:** The hybrid model outperformed individual models, illustrating the benefits of integrating multiple algorithms for enhanced accuracy.

2.2 TECHNIQUES AND MODELS: SUMMARY OF PREDICTIVE MODELLING TECHNIQUES COMMONLY USED IN CHURN PREDICTION

Churn prediction models are designed to identify customers likely to leave a service provider. Various machine learning techniques are employed, each with its strengths and limitations. The Random Forest algorithm is among the most effective and widely used techniques. Here is a summary of key techniques:

1. **Regression:**
    - **Description:** A statistical method for analyzing the relationship between one or more independent variables and an outcome.
    - **Strengths:** Simple to implement and interpret; suitable for binary classification.
    - **Weaknesses:** Assumes a linear relationship between variables and the outcome, which may not capture complex patterns.

2. **Decision Trees:**
    - **Description:** A model that makes decisions based on a series of rules derived from data features.
    - **Strengths:** Easy to visualize and interpret; handles both numerical and categorical data.
    - **Weaknesses:** Prone to overfitting and sensitive to small changes in the data.
3. **Random Forest:**
    - **Description:** An ensemble learning method that builds multiple decision trees during training and outputs the mode of the classes (for classification) or mean prediction (for regression) of the individual trees.
    - **Strengths:** Reduces overfitting by averaging multiple decision trees; handles large datasets with high dimensionality; robust to noise and outliers.
    - **Weaknesses:** Can be computationally intensive and less interpretable than single decision trees.
4. **Support Vector Machines (SVM):**
    - **Description:** A supervised learning model that classifies data by finding the hyperplane that maximizes the margin between classes.
    - **Strengths:** Effective in high-dimensional spaces; versatile with different kernel functions.
    - **Weaknesses:** Requires careful parameter tuning; computationally expensive for large datasets.
5. **Neural Networks:**
    - **Description:** Models inspired by the human brain's structure, capable of capturing complex patterns through multiple layers of interconnected neurons.
    - **Strengths:** Excellent at capturing nonlinear relationships; scalable to large datasets.
    - **Weaknesses:** Requires large amounts of data and computational resources; prone to overfitting without proper regularization.
6. **Gradient Boosting Machines (GBM):**
    - **Description:** An ensemble technique that builds models sequentially, with each new model correcting errors made by previous ones.
    - **Strengths:** High predictive accuracy; effective for both regression and classification problems.
    - **Weaknesses:** Computationally expensive; sensitive to overfitting if not tuned correctly.

Random Forest for Telecom Customer Loss Prediction: Why Random Forest?

Random Forest is particularly effective for telecom customer churn prediction due to its capacity to handle large and complex datasets with numerous features. Its ability to mitigate overfitting by averaging results from multiple trees, along with its robustness to noise and outliers, makes it a preferred choice for modeling customer churn, which often involves high-dimensional and noisy data.

Implementation Steps:

1. **Data Preprocessing:**
    - Collect and clean data from various sources (e.g., CRM systems, billing records, support logs).
    - Handle missing values and outliers.
    - Encode categorical variables and normalize numerical features.
2. **Feature Engineering:**
    - Identify and create relevant features such as customer tenure, average call duration, and frequency of service usage.
    - Use domain knowledge to derive additional features that may influence churn.
3. **Model Training:**
    - Split the data into training and testing sets.
    - Train the Random Forest model on the training data, tuning hyperparameters such as the number of trees, maximum depth, and minimum samples split.
4. **Model Evaluation:**
    - Evaluate model performance using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.
    - Compare results with baseline models (e.g., logistic regression) and other advanced models (e.g., GBM, neural networks).
5. **Deployment:**
    - Deploy the model in a production environment for real-time or batch predictions.
    - Integrate the model with CRM systems to automate retention strategies for at-risk customers.

CHAPTER 3

DATA COLLECTION AND MANAGEMENT

Data Sources: Description of Data Sources

In a telecom customer loss prediction project, accurate and comprehensive data from various sources is crucial for developing an effective predictive model. The primary data sources typically include:

1. **Customer Relationship Management (CRM) Systems:**
    - **Content:** Customer demographics, contact information, customer tenure, contract details, and interaction history.
    - **Significance:** Provides a detailed view of the customer's profile and engagement with the company, which is vital for understanding customer behavior and potential churn.
2. **Billing Systems:**
    - **Content:** Payment history, billing cycles, outstanding balances, late payments, and transaction records.
    - **Significance:** Financial behavior and payment history are strong indicators of customer satisfaction and potential churn, offering insights into the financial aspects of customer relationships.
3. **Customer Support Logs:**
    - **Content:** Records of customer inquiries, complaints, service requests, and resolution times.
    - **Significance:** A high frequency of complaints and unresolved issues can indicate dissatisfaction and an increased risk of churn, highlighting areas needing improvement.
4. **Usage Data:**
    - **Content:** Call detail records (CDRs), data usage, text message volumes, service plan details, and usage patterns.
    - **Significance:** Usage behavior helps identify trends and anomalies that may signal churn risk, providing insights into how customers interact with the services.
5. **Marketing and Campaign Data:**
    - **Content:** Records of marketing campaigns, promotional offers, customer responses, and engagement rates.

- **Significance:** Understanding the effectiveness of marketing efforts and their influence on customer retention can guide future strategies and interventions.

6. **Social Media and Online Reviews:**
    - **Content:** Customer reviews, social media mentions, and sentiment analysis.
    - **Significance:** Offers insights into customer sentiment and public perception of the company, helping to gauge overall satisfaction and areas for improvement.
7. **Network Data:**
    - **Content:** Service quality metrics, network usage, and outage records.
    - **Significance:** Service reliability and quality are critical to customer satisfaction and retention, with network issues potentially driving customers to seek alternative providers.

Data Integration: How Data from Different Sources Was Combined

Integrating data from multiple sources is a crucial step in creating a unified dataset for churn prediction. This process involves:

1. **Data Extraction:**
    - Extract data from various sources using APIs, database queries, or file exports.
    - Ensure data is collected at regular intervals to maintain consistency and timeliness.
2. **Data Cleaning:**
    - Remove duplicate records across different datasets to prevent redundancy.
    - Standardize data formats (e.g., date formats, currency units) to ensure uniformity.
3. **Data Matching:**
    - Match records from different sources based on unique identifiers such as customer ID, phone number, or email address.
    - Resolve discrepancies in matching records to ensure accuracy and completeness.
4. **Data Merging:**
    - Merge data from different sources into a single, unified dataset.
    - Employ relational database techniques or data warehousing solutions to efficiently combine large datasets.
5. **Schema Alignment:**
    - Ensure the combined dataset has a consistent schema with standardized column names and data types.

- Document the schema to facilitate future data management and analysis.

6. **Handling Missing Values:**
   - Identify and address missing values in the dataset.
   - Use techniques such as imputation, default values, or exclusion of records with excessive missing data to manage gaps.
7. **Data Transformation:**
   - Transform data into a suitable format for analysis, including normalizing numerical values and encoding categorical variables.
   - Create new features from existing data to enhance the predictive power of the model.
8. **Data Storage:**
   - Store the integrated dataset in a secure, scalable database or data warehouse.
   - Ensure regular backups to protect against data loss.

Data Preprocessing: Methods Used for Cleaning, Handling Missing Values, Outlier Detection, etc.

Data preprocessing is essential for preparing the dataset for model training and ensuring the accuracy and reliability of predictions. Key preprocessing steps include:

1. **Data Cleaning:**
   - **Duplicate Removal:** Identify and remove duplicate records to avoid model bias and ensure data integrity.
   - **Consistency Checks:** Verify consistency across different fields (e.g., units, formats) to maintain data quality.
2. **Handling Missing Values:**
   - **Detection:** Identify missing values using descriptive statistics or data visualization techniques.
   - **Imputation:** Fill missing values using methods such as mean, median, mode imputation, or advanced techniques like k-nearest neighbors (KNN) imputation.
   - **Exclusion:** Exclude records with excessive missing values if imputation is not feasible.
3. **Outlier Detection and Handling:**
   - **Detection:** Identify outliers using statistical methods (e.g., z-scores, interquartile range) or visualization techniques (e.g., box plots).
   - **Handling:** Decide on a strategy for handling outliers, such as removal, transformation, or capping.

4. **Normalization and Scaling:**
    - **Normalization:** Normalize numerical features to a common scale (e.g., 0 to 1) to ensure they contribute equally to the model.
    - **Standardization:** Standardize features to have a mean of zero and a standard deviation of one, improving model performance.
5. **Encoding Categorical Variables:**
    - **Label Encoding:** Convert categorical variables into numerical labels when there is an ordinal relationship between categories.
    - **One-Hot Encoding:** Create binary columns for each category when there is no ordinal relationship, avoiding the imposition of unintended hierarchies.

CHAPTER 4

FEATURE ENGINEERING

Feature Selection: Criteria for Selecting Relevant Features

Feature selection is a crucial step in developing an effective churn prediction model. The objective is to identify and retain the most relevant features that significantly contribute to predicting customer churn. Key criteria for selecting relevant features include:

1. **Domain Knowledge:**
    - **Description:** Leverage insights from telecom industry experts to identify features that are known to impact customer churn.
    - **Examples:** Features such as call quality, billing issues, and customer service interactions are often influential.
2. **Correlation Analysis:**
    - **Description:** Assess the correlation between each feature and the target variable (churn) to identify features with strong relationships.
    - **Methods:** Use Pearson's correlation coefficient for continuous features and Cramér's V for categorical features.
3. **Variance Threshold:**
    - **Description:** Eliminate features with low variance, as these features contribute minimal information about the target variable.
    - **Application:** Features with very little variation are less likely to impact the model's predictive performance.
4. **Mutual Information:**
    - **Description:** Measure the mutual information between features and the target variable to capture non-linear relationships.
    - **Purpose:** This method helps identify features that have a significant relationship with the target variable, even if it's not linear.
5. **Feature Importance from Models:**
    - **Description:** Utilize tree-based models, such as Random Forest and Gradient Boosting, to obtain feature importance scores.
    - **Application:** Select features with high importance scores as they are more influential in predicting churn.

6. **Recursive Feature Elimination (RFE):**
    - **Description:** RFE is used to recursively remove the least important features and build the model iteratively.
    - **Purpose:** This technique retains only the most significant features by evaluating model performance at each step.
7. **Backward and Forward Selection:**
    - **Description:** Implement stepwise regression techniques to add or remove features based on their contribution to model performance.
    - **Methods:** Backward selection starts with all features and removes the least significant ones, while forward selection begins with no features and adds the most significant ones.
8. **Stability Selection:**
    - **Description:** Apply stability selection methods to ensure that selected features remain consistent across different data samples.
    - **Purpose:** This approach helps in identifying robust features that are stable and reliable across various datasets.

Examples of Important Features in Telecom Churn Prediction:

- **Customer Tenure:** The length of time the customer has been with the company. Longer tenure may correlate with lower churn risk.
- **Billing Issues:** Frequency and severity of billing disputes or late payments. Issues with billing can be a significant indicator of potential churn.
- **Usage Patterns:** Average call duration, data usage, and text message volumes. These patterns can reveal changes in customer behavior that may indicate churn.
- **Customer Support Interactions:** The number and type of interactions with customer support, particularly unresolved issues. Frequent or unresolved issues may increase churn risk.
- **Service Quality:** Metrics such as call drop rate, network reliability, and service outages. Poor service quality can drive customers to seek alternatives.
- **Marketing Engagement:** Response to promotional offers and participation in marketing campaigns. Engaged customers may have a lower churn risk.
- **Sentiment Scores:** Sentiment analysis of customer reviews and support interactions. Positive or negative sentiments can influence churn probability.
- **Contract Details:** Type of contract, contract length, and time remaining on the contract. Contractual aspects can impact customer retention and churn risk.

CHAPTER 5

PREDICTIVE MODELING

Model Selection: Justification for Choosing Specific Models

In telecom customer loss prediction, choosing the right predictive models is essential for achieving accurate and actionable insights. The selection process considers data characteristics, the complexity of relationships between variables, and model interpretability. Here are some commonly used models and the rationale behind their selection:

1. **Logistic Regression:**
    - **Justification:**
        - Simple and interpretable model that provides probabilistic outputs.
        - Suitable for binary classification tasks like churn prediction.
        - Effective when the relationship between independent variables and the target variable is approximately linear.
    - **Limitations:**
        - May not capture complex, non-linear relationships in the data.
2. **Decision Trees:**
    - **Justification:**
        - Easy to understand and visualize.
        - Can handle both numerical and categorical data.
        - Capable of capturing non-linear relationships.
    - **Limitations:**
        - Prone to overfitting, especially with deep trees.
        - Sensitive to small variations in the data.
3. **Random Forest:**
    - **Justification:**
        - An ensemble learning method that reduces overfitting by averaging multiple decision trees.
        - Robust to noise and can handle large datasets with many features.
        - Provides feature importance scores, aiding in feature selection and interpretation.

- **Limitations:**
  - Computationally intensive, especially with a large number of trees.

4. **Gradient Boosting Machines (GBM):**
   - **Justification:**
     - Builds models sequentially, with each new model correcting errors made by the previous ones.
     - High predictive accuracy and can capture complex patterns in the data.
   - **Limitations:**
     - More computationally expensive and requires careful tuning to avoid overfitting.
5. **Support Vector Machines (SVM):**
   - **Justification:**
     - Effective in high-dimensional spaces and when the classes are not linearly separable.
     - Uses kernel functions to capture non-linear relationships.
   - **Limitations:**
     - Not suitable for very large datasets due to high computational cost.
     - Requires careful tuning of hyperparameters.
6. **K-Nearest Neighbours (KNN):**
   - **Justification:**
     - Simple and intuitive model that makes predictions based on the closest training examples.
     - Effective when the decision boundary is highly non-linear.
   - **Limitations:**
     - Computationally expensive during prediction time.
     - Sensitive to the choice of k and the distance metric.

Model Training: Steps Taken to Train the Models

Training a predictive model for telecom customer loss prediction involves several key steps:

1. **Data Preprocessing:**
   - **Cleaning:** Handle missing values, remove duplicates, and address outliers.
   - **Normalization/Standardization:** Normalize or standardize numerical features to ensure equal contribution to the model.

- **Encoding:** Encode categorical variables using techniques like one-hot encoding or label encoding.
- **Splitting:** Divide the data into training and testing sets (e.g., 80% training, 20% testing).

2. **Feature Engineering:**
   - **Selection:** Choose relevant features based on domain knowledge, correlation analysis, and feature importance scores.
   - **Creation:** Create new features to capture important patterns and relationships in the data.
   - **Scaling:** Perform feature scaling to ensure that all features are on a similar scale.
3. **Model Selection:**
   - **Diversity:** Choose a diverse set of models (e.g., logistic regression, decision trees, random forest, GBM) for initial training.
   - **Requirements:** Consider specific requirements of the churn prediction task, such as the need for interpretability or the ability to handle large datasets.
4. **Model Training:**
   - **Initialization:** Start with default hyperparameters for each model.
   - **Training:** Train the models on the training dataset to capture the relationship between features and the target variable.

Model Validation: Techniques Used to Validate Model Performance

Validating the performance of a predictive model is essential to ensure its reliability and generalizability. Common validation techniques include:

1. **Train-Test Split:**
   - **Description:** Split the dataset into training and testing sets to evaluate the model on unseen data.
   - **Consideration:** Ensure the split is random but stratified to maintain the proportion of churners and non-churners in both sets.
2. **Cross-Validation:**
   - **Description:** Implement k-fold cross-validation to divide the data into k subsets (folds).
   - **Procedure:** Train the model k times, each time using a different fold as the test set and the remaining folds as the training set.
   - **Output:** Average the performance metrics across all folds to obtain a robust estimate of model performance.

3. **Performance Metrics:**
    - **Accuracy:** Proportion of correctly predicted instances out of the total instances.
    - **Precision:** Proportion of true positive predictions among all positive predictions.
    - **Recall (Sensitivity):** Proportion of true positive predictions among all actual positives.
    - **F1-Score:** Harmonic mean of precision and recall, providing a balance between the two.
    - **AUC-ROC (Area Under the Receiver Operating Characteristic Curve):** Measures the ability of the model to distinguish between churners and non-churners.
    - **Confusion Matrix:** Provides a detailed breakdown of true positives, true negatives, false positives, and false negatives.
4. **Hyperparameter Tuning:**
    - **Techniques:** Use techniques like grid search or random search to find the optimal hyperparameters for each model.
    - **Validation:** Evaluate the models using cross-validation during the hyperparameter tuning process.

CHAPTER 6

USER INTERFACE

Dashboard Design: Description of the Dashboard for Visualizing Churn Predictions

A well-designed dashboard is crucial for visualizing churn predictions and providing actionable insights to stakeholders. The dashboard should be intuitive, interactive, and provide comprehensive information to help telecom companies identify at-risk customers and take proactive measures. Key components of the dashboard for telecom customer loss prediction include:

1. **Overview Section:**
    - **Churn Rate:** A visual representation (e.g., gauge chart) of the overall churn rate for the selected period.
    - **Total Customers:** The current total number of customers and the number of churned customers.
    - **Churn Trend:** A line chart showing the trend of churn rates over time, helping to identify seasonal patterns or the impact of recent changes.
2. **Customer Segmentation:**
    - **Churn by Segment:** Bar charts or pie charts showing churn rates across different customer segments (e.g., by demographic, service plan, region).
    - **High-Risk Segments:** Highlight segments with the highest churn rates for targeted interventions.
3. **Customer Details:**
    - **Customer List:** A table listing individual customers with their churn probability scores, customer ID, name, service plan, and other relevant details.
    - **Drill-Down Capabilities:** Ability to click on a customer to view detailed information, including usage patterns, billing history, support interactions, and service quality metrics.
4. **Alerts and Notifications:**
    - **At-Risk Alerts:** Automated alerts for customers with high churn probability, enabling timely interventions.
    - **Service Quality Alerts:** Notifications for customers experiencing service issues that may contribute to churn.
5. **Actionable Recommendations:**
    - **Retention Strategies:** Suggested actions for retaining high-risk customers, such as targeted offers, personalized communications, or loyalty programs.

- **Effectiveness Tracking:** Monitor the outcomes of implemented retention strategies and adjust based on effectiveness.

6. **Customizable Views:**
    - **Filters:** Ability to filter data by date range, customer segment, region, service plan, and other criteria.
    - **User-Specific Dashboards:** Personalized views for different roles (e.g., customer service, marketing, management) to ensure relevant information is readily accessible.

Conclusion

The telecom customer loss prediction project effectively addresses the critical issue of customer churn by leveraging advanced predictive modeling techniques. By analyzing data from CRM systems, billing records, and customer support logs, the project integrates and preprocesses information to ensure accuracy and completeness.

Key components include:

- **Data-Driven Insights:** Advanced models like Random Forests and Gradient Boosting Machines provide robust and accurate churn predictions.
- **User-Friendly Dashboard:** A well-designed dashboard and reporting tools offer actionable insights, enabling timely interventions to retain customers.
- **Comprehensive Training:** Stakeholders are equipped with the knowledge to use the system effectively through tailored training programs.

This approach not only predicts which customers are at risk of churning but also offers targeted strategies for retention, ultimately helping telecom companies reduce churn rates, improve customer satisfaction, and enhance business performance.

Program Code:

```
#import the required libraries

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.ticker as mtick

import matplotlib.pyplot as plt

%matplotlib inline


# Load data

telco_base_data = pd.read_csv('WA_Fn-UseC_-Telco-Customer-Churn.csv')

telco_base_data.head()
```

```
telco_base_data.shape
telco_base_data.columns.values

# Checking the data types of all columns
telco_base_data.dtypes

# Checking descriptive statistics of numeric variables
telco_base_data.describe()
telco_base_data['Churn'].value_counts().plot(kind='barh', figsize=(8, 6))
plt.xlabel("Count", labelpad=14)
plt.ylabel("Target Variable", labelpad=14)
plt.title("Count of TARGET Variable per category", y=1.02)

# Concise summary of the dataframe
telco_base_data.info(verbose=True)

# Checking for missing values
missing = pd.DataFrame((telco_base_data.isnull().sum()) * 100 / telco_base_data.shape[0]).reset_index()
plt.figure(figsize=(16, 5))
ax = sns.pointplot(x='index', y=0, data=missing) # Specify 'x' and 'y' explicitly
plt.xticks(rotation=90, fontsize=7)
plt.title("Percentage of Missing Values")
plt.ylabel("PERCENTAGE")
plt.show()

# Copy data and convert TotalCharges to numeric
telco_data = telco_base_data.copy()
telco_data.TotalCharges = pd.to_numeric(telco_data.TotalCharges, errors='coerce')
telco_data.isnull().sum()
```

```
# Removing missing values
telco_data.dropna(how='any', inplace=True)


# Get the max tenure and group tenure in bins of 12 months
print(telco_data['tenure'].max()) # 72
labels = ["{0}-{1}".format(i, i+11) for i in range(1, 72, 12)]
telco_data['tenure_group'] = pd.cut(telco_data.tenure, range(1, 80, 12), right=False, labels=labels)
# Drop columns
telco_data.drop(columns=['customerID', 'tenure'], axis=1, inplace=True)
telco_data.head()


# Plotting predictors
for i, predictor in enumerate(telco_data.drop(columns=['Churn', 'TotalCharges', 'MonthlyCharges'])):
    plt.figure(i)
    sns.countplot(data=telco_data, x=predictor, hue='Churn')


# Convert Churn to binary
telco_data['Churn'] = np.where(telco_data.Churn == 'Yes', 1, 0)
telco_data.head()


# Creating dummy variables
telco_data_dummies = pd.get_dummies(telco_data)
telco_data_dummies.head()


# Plot Monthly Charges vs Total Charges
sns.lmplot(data=telco_data_dummies, x='MonthlyCharges', y='TotalCharges', fit_reg=False)


# Density plots
```

```
Mth = sns.kdeplot(telco_data_dummies.MonthlyCharges[(telco_data_dummies["Churn"] == 0)], color="Red", shade=True)

Mth = sns.kdeplot(telco_data_dummies.MonthlyCharges[(telco_data_dummies["Churn"] == 1)], ax=Mth, color="Blue", shade=True)

Mth.legend(["No Churn", "Churn"], loc='upper right')

Mth.set_ylabel('Density')

Mth.set_xlabel('Monthly Charges')

Mth.set_title('Monthly Charges by Churn')


Tot = sns.kdeplot(telco_data_dummies.TotalCharges[(telco_data_dummies["Churn"] == 0)], color="Red", shade=True)

Tot = sns.kdeplot(telco_data_dummies.TotalCharges[(telco_data_dummies["Churn"] == 1)], ax=Tot, color="Blue", shade=True)

Tot.legend(["No Churn", "Churn"], loc='upper right')

Tot.set_ylabel('Density')

Tot.set_xlabel('Total Charges')

Tot.set_title('Total Charges by Churn')


# Correlation matrix and heatmap

plt.figure(figsize=(20, 8))

telco_data_dummies.corr()['Churn'].sort_values(ascending=False).plot(kind='bar')

plt.figure(figsize=(12, 12))

sns.heatmap(telco_data_dummies.corr(), cmap="Paired")


# Separate datasets for churn and no churn

new_df1_target0 = telco_data.loc[telco_data["Churn"] == 0]

new_df1_target1 = telco_data.loc[telco_data["Churn"] == 1]


# Function for plotting

def uniplot(df, col, title, hue=None):

    sns.set_style('whitegrid')

    sns.set_context('talk')
```

```
    plt.rcParams["axes.labelsize"] = 20
    plt.rcParams['axes.titlesize'] = 22
    plt.rcParams['axes.titlepad'] = 3
    temp = pd.Series(data=hue)
    fig, ax = plt.subplots()
    width = len(df[col].unique()) + 7 + 4 * len(temp.unique())
    fig.set_size_inches(width, 8)
    plt.xticks(rotation=45)
    plt.yscale('log')
    plt.title(title)
    ax = sns.countplot(data=df, x=col, order=df[col].value_counts().index, hue=hue, palette='bright')
    plt.show()


# Create a dictionary to map inputs
input_dict = {
    'SeniorCitizen': [SeniorCitizen],
    'MonthlyCharges': [MonthlyCharges],
    'TotalCharges': [TotalCharges],
    'gender_Female': [1 if gender == "Female" else 0],
    'gender_Male': [1 if gender == "Male" else 0],
    'Partner_No': [1 if Partner == "No" else 0],
    'Partner_Yes': [1 if Partner == "Yes" else 0],
    'Dependents_No': [1 if Dependents == "No" else 0],
    'Dependents_Yes': [1 if Dependents == "Yes" else 0],
    'PhoneService_No': [1 if PhoneService == "No" else 0],
    'PhoneService_Yes': [1 if PhoneService == "Yes" else 0],
    'MultipleLines_No': [1 if MultipleLines == "No" else 0],
    'MultipleLines_No phone service': [1 if MultipleLines == "No phone service" else 0],
    'MultipleLines_Yes': [1 if MultipleLines == "Yes" else 0],
    'InternetService_DSL': [1 if InternetService == "DSL" else 0],
```

```
    'InternetService_Fiber optic': [1 if InternetService == "Fiber optic" else 0],
    'InternetService_No': [1 if InternetService == "No" else 0],
    'OnlineSecurity_No': [1 if OnlineSecurity == "No" else 0],
    'OnlineSecurity_No internet service': [1 if OnlineSecurity == "No internet service" else 0],
    'OnlineSecurity_Yes': [1 if OnlineSecurity == "Yes" else 0],
    'OnlineBackup_No': [1 if OnlineBackup == "No" else 0],
    'OnlineBackup_No internet service': [1 if OnlineBackup == "No internet service" else 0],
    'OnlineBackup_Yes': [1 if OnlineBackup == "Yes" else 0],
    'DeviceProtection_No': [1 if DeviceProtection == "No" else 0],
    'DeviceProtection_No internet service': [1 if DeviceProtection == "No internet service" else
0],
    'DeviceProtection_Yes': [1 if DeviceProtection == "Yes" else 0],
    'TechSupport_No': [1 if TechSupport == "No" else 0],
    'TechSupport_No internet service': [1 if TechSupport == "No internet service" else 0],
    'TechSupport_Yes': [1 if TechSupport == "Yes" else 0],
    'StreamingTV_No': [1 if StreamingTV == "No" else 0],
    'StreamingTV_No internet service': [1 if StreamingTV == "No internet service" else 0],
    'StreamingTV_Yes': [1 if StreamingTV == "Yes" else 0],
    'StreamingMovies_No': [1 if StreamingMovies == "No" else 0],
    'StreamingMovies_No internet service': [1 if StreamingMovies == "No internet service" else
0],
    'StreamingMovies_Yes': [1 if StreamingMovies == "Yes" else 0],
    'Contract_Month-to-month': [1 if Contract == "Month-to-month" else 0],
    'Contract_One year': [1 if Contract == "One year" else 0],
    'Contract_Two year': [1 if Contract == "Two year" else 0],
    'PaperlessBilling_No': [1 if PaperlessBilling == "No" else 0],
    'PaperlessBilling_Yes': [1 if PaperlessBilling == "Yes" else 0],
    'PaymentMethod_Bank transfer (automatic)': [1 if PaymentMethod == "Bank transfer
(automatic)" else 0],
    'PaymentMethod_Credit card (automatic)': [1 if PaymentMethod == "Credit card
(automatic)" else 0]
}
```

REFERENCES

- **Towards Data Science - Customer Churn Prediction:**

  https://www.google.com/search?q=Towards+Data+Science+Customer+Churn+Prediction
  This article provides a detailed overview of customer churn prediction, including methodologies, data preparation, and model implementation using Python.

- **Kaggle - Telco Customer Churn Dataset:** Link to Dataset This page provides access to a popular dataset for customer churn analysis in the telecommunications industry, along with kernels (code examples) for data analysis and model building.

  https://www.google.com/search?q=Kaggle+Telco+Customer+Churn+Datset

- **IBM Developer - Customer Churn Analysis:**
  https://www.google.com/search?q=IBM+Developer+Customer+Churn+Analysis
  This article discusses the importance of customer churn prediction and provides insights into how IBM's solutions can be used to address churn prediction challenges.

- **Scikit-learn Documentation:**

  https://www.google.com/search?q=Scikitlearn+Documentation+Logistic+Regression
  The official documentation for the logistic regression module in scikit-learn, a widely used Python library for machine learning. It includes explanations, usage examples, and parameters.

- **Machine Learning Mastery - How to Develop a Customer Churn Prediction Model:** Link to Tutorial This tutorial walks through the process of developing a customer churn prediction model in Python, covering data preparation, model training, and evaluation.
  https://www.google.com/search?q=Machine+Learning+Mastery+How+to+Develop+a+Customer+Churn+Prediction+Model