Executive Summary

This analysis provides actionable insights into customer churn prediction and strategies for improving retention. Through **Data Analysis** and **Exploratory Data Analysis (EDA)**, along with **statistical modeling**, feature analysis, and customer behavior exploration, several key findings emerged that can drive business decisions.

Data Analysis and Exploratory Data Analysis (EDA):

1. **Data Preprocessing**:
    - The dataset consists of various features such as **tenure**, **contract type**, **monthly charges**, **total charges**, and **service usage**.
    - We performed **data cleaning** to handle missing values, convert categorical variables into numerical representations, and ensure the data was ready for modeling.
2. **Visualizations**:
    - **Pie chart** analysis showed that **26.54% of customers** have churned, representing **1,869 customers**.
    - **Box plots and histograms** revealed that **long-tenured customers** (those with higher tenure) tend to have higher **total charges**, and customers on **long-term contracts** (1- or 2-year contracts) are less likely to churn.
    - A **heatmap** analysis uncovered strong correlations between features like **tenure** and **total charges (0.83)**, and **tenure** and **contract type (0.67)**, indicating that customers who stay longer with the company tend to accumulate higher charges and are more likely to be on long-term contracts.
3. **Key Insights from EDA**:
    - **Customers with partners or dependents** tend to have lower churn rates.
    - **Paperless billing customers** show a slightly higher churn rate.
    - **Service Usage**: Customers using services like **PhoneService**, **InternetService (DSL)**, and **OnlineSecurity** are less likely to churn, whereas those not using **OnlineBackup**, **TechSupport**, or **StreamingTV** are more likely to churn.
    - **Gender analysis**: Churning is **not gender-specific**, suggesting that other factors (like service usage and contract type) have a more critical role in churn prediction.
4. **Churn Distribution**:
    - A breakdown of churn shows that the churn rate is fairly evenly distributed across different demographic groups, with **26.54% of customers** churning out, emphasizing the importance of predicting and reducing churn.

Key Findings:

1. **Customer Demographics and Behavior**:
    - **Customers with partners or dependents** are significantly less likely to churn.
    - **Customers using paperless billing** have a slightly higher churn rate, possibly indicating disengagement.
    - **Service Usage**:
        - Customers who have **PhoneService**, **InternetService**, and **OnlineSecurity** activated are less likely to churn.
        - High churn rates are associated with customers who do not use services like **OnlineBackup**, **TechSupport**, or **StreamingTV**.
2. **Correlation Insights**:
    - **Tenure** and **TotalCharges** (0.83) show a strong positive correlation, indicating that longer-tenured customers tend to accumulate higher charges.
    - **Tenure** and **Contract** (0.67) suggest that long-tenured customers are more likely to have long-term contracts, reinforcing the importance of contract type in retention.

    - Other correlations include **Partner** with **Dependents** (0.45), **MultipleLines** with **MonthlyCharges** and **TotalCharges** (0.43-0.45), and **TechSupport** with **Contract** (0.43).
3. **Churn Risk Factors**:
    - **Month-to-month contract holders** are more likely to churn compared to customers on **1- or 2-year contracts**.
    - **Long-tenured customers** are less likely to churn, whereas newer customers are more prone to churn.
4. **Churn Data**:
    - From the pie chart, **26.54% of customers** have churned, representing **1,869 customers**.
    - Churning is **not gender-specific**, indicating that churn is more related to other factors like service usage and contract type.

Model Performance:

- **Logistic Regression** and **Random Forest** models were evaluated for churn prediction:
    - **Logistic Regression** showed an accuracy of **81.3%**, with a **recall** of **83%** for churn prediction. This makes Logistic Regression more suitable when the goal is to minimize false negatives (ensuring most churners are detected). However, it has a **lower precision** (52%), meaning it misclassifies some non-churners as churners.
    - **Random Forest** demonstrated an accuracy of **79.2%** with a **precision** of **57%**, making it more balanced but sacrificing some recall (64%). This model is better at avoiding false churners but may miss some churn cases.

Feature Importance:

- **High Impact Features**:
    - **MonthlyCharges**, **TotalCharges**, **Tenure**, and **Contract** are the most significant predictors of churn.
    - **Customers with higher charges** or **paperless billing** are more likely to churn, while **long-tenure** customers and those with **long-term contracts** are less likely to churn.
- **Moderate Impact**:
    - **OnlineSecurity**, **TechSupport**, **PaymentMethod**, and **OnlineBackup** show moderate effects on churn, with these services contributing to customer retention.
- **Low Impact Features**:
    - **Partner**, **MultipleLines**, and **StreamingTV** have minimal influence on churn and should not be prioritized in retention efforts.

Strategic Recommendations:

1. **Targeting High-Risk Customers**:
    - Focus retention efforts on customers with **higher monthly charges** or those using **paperless billing**, as these features are correlated with a higher likelihood of churn.
    - Customers on **month-to-month contracts** should be targeted with retention strategies, as they are more likely to churn.
2. **Enhancing Service Engagement**:
    - Ensure that services such as **PhoneService**, **InternetService**, and **OnlineSecurity** are widely available and promoted to reduce churn.
    - Improve **TechSupport** and **OnlineBackup** availability, as these services help with customer retention.

3. **Long-Term Customer Retention**:
    - Use **Tenure** and **Contract** features to offer incentives for customers to stay longer or transition to long-term contracts, as longer-tenured customers are less likely to churn.
4. **Model Optimization**:
    - **Logistic Regression** is preferable for identifying at-risk customers due to its high recall, while **Random Forest** offers a more balanced approach. Further improvements can be made through **hyperparameter tuning** to optimize model performance.

What I Have Learned:

Customer Behavior Analysis:

In this analysis, I leveraged Python and its powerful data analysis libraries (such as Pandas and NumPy) to deeply explore customer behavior and churn patterns. Through **Exploratory Data Analysis (EDA)**, I identified that features like **Tenure**, **Contract type**, and **Service usage** (e.g., PhoneService, InternetService) play a crucial role in determining churn. Customers with **longer tenures**, **long-term contracts**, and essential services are less likely to churn, whereas **paperless billing** and **higher monthly charges** correlate with a higher likelihood of churn. My use of visualizations with **Matplotlib** and **Seaborn** helped uncover these relationships clearly.

Modeling Techniques & Experimentation:

Using Python's **scikit-learn** library, I applied various **machine learning models** (Logistic Regression and Random Forest) to predict customer churn. I fine-tuned the models through **GridSearchCV** to find the best hyperparameters, and implemented cross-validation to ensure model robustness. Logistic Regression showed strong performance in churn detection, excelling at **high recall** (identifying most churners), while Random Forest was effective in **balancing precision** with recall. This experience helped me understand the strengths and limitations of different models and their applications.

Feature Importance & Analysis:

I utilized **feature importance techniques** to evaluate which variables most influenced churn. By analyzing **Logistic Regression coefficients** and **Random Forest feature importances**, I was able to identify key drivers of churn, such as **MonthlyCharges**, **TotalCharges**, **Tenure**, and **Contract**. Understanding these features allowed me to refine predictions and make data-driven recommendations for improving customer retention. I gained hands-on experience with Python's **pandas** for feature extraction and transformation, as well as visualizing relationships using **seaborn**.

Data Balancing with SMOTE:

I applied **SMOTE (Synthetic Minority Over-sampling Technique)** to address class imbalance in churn prediction. Using **imbalanced-learn** in Python, I successfully balanced the churn cases, ensuring that the models did not overlook the minority churn class. This led to better model performance and more accurate predictions, underscoring the importance of data balancing in real-world machine learning problems.

Strategic Application & Business Impact:

By combining advanced **data analysis** and **machine learning**, I developed actionable insights that can directly influence business strategies. Understanding churn drivers such as **charges, tenure, and contract type** enables businesses to focus on high-risk customer segments and tailor retention strategies. This project showcased my ability to not only perform data analysis but also derive actionable insights that improve business outcomes.