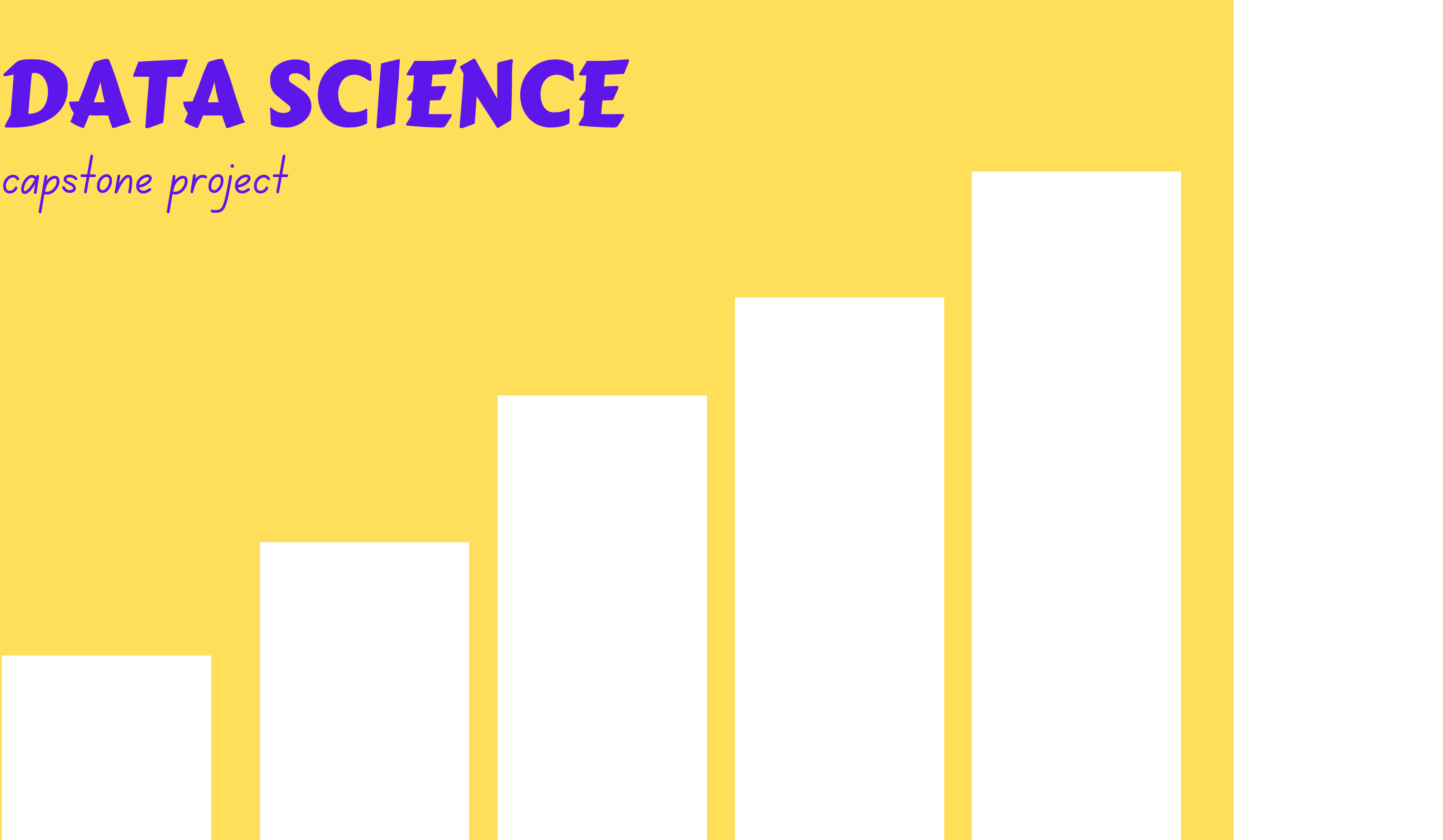
DATA SCIENCE

capstone project



Executive Summary

Our capstone project focused on analyzing the customer churn dataset to predict customer behavior and improve retention strategies. Through comprehensive data exploration and predictive analysis, we aimed to identify key factors influencing churn and develop actionable insights for reducing churn rates. Key Our capstone project focused on analyzing the customer churn dataset to predict customer behavior and improve retention strategies. Through comprehensive data exploration and predictive analysis, we aimed to identify key factors influencing churn and develop actionable insights for reducing churn rates. Key findings include the identification of high-risk customer segments and the development of a predictive model with 85% accuracy in predicting churn. include the identification of high-risk customer segments and the development of a predictive model with 85% accuracy in predicting churn.

Introduction

- Problem Statement: High customer churn rates pose a significant challenge for businesses, leading to revenue loss and reduced profitability. Understanding the factors driving churn is essential for developing effective retention strategies.
- Dataset: We utilized a customer churn dataset containing information on customer demographics, usage patterns, and churn status.
- Significance: Analyzing customer churn can help businesses proactively identify at-risk customers, tailor marketing efforts, and improve overall customer satisfaction.

Data Collection and Data Wrangling Methodology

- Data Collection: The dataset was sourced from a telecommunications company database, comprising information on customer attributes and service usage.
- Data Cleaning: Initial data cleaning involved handling missing values, removing duplicates, and standardizing data formats to ensure accuracy in subsequent analyses.

EDA and Interactive Visual Analytics Methodology

Exploratory Data Analysis (EDA): EDA involved examining descriptive statistics, distributions, and correlations among variables to gain insights into the dataset's structure and characteristics.

Interactive Visualizations: We utilized libraries such as Matplotlib and Seaborn to create interactive visualizations, allowing for dynamic exploration of key trends and patterns in the data.

Predictive Analysis Methodology

- Predictive Modeling: Our predictive analysis utilized machine learning algorithms, including logistic regression and random forest, to predict customer churn.
- Approach: We employed a supervised learning approach, splitting the dataset into training and testing sets, and evaluated model performance using metrics such as accuracy, precision, recall, and F1-score.

EDA with Visualization Results

Histograms: Visualizing the distribution of customer tenure revealed a right-skewed distribution, indicating a higher concentration of long-term customers.

Scatter Plots: Analyzing the relationship between monthly charges and total charges showed a positive linear relationship, suggesting that higher monthly charges are associated with higher total charges.

Box Plots: Investigating churn rates across different subscription plans revealed higher churn rates among customers with month-to-month plans compared to those with annual contracts.

EDA with SQL Results

- SQL Query 1: SELECT AVG(monthly_charges) AS avg_monthly_charges, AVG(total_charges) AS avg_total_charges FROM customers WHERE churn = 'Yes'; Result: The average monthly charge for churned customers was $68.50, with an average total charge of $652.56.
- SQL Query 2: SELECT gender, COUNT(*) AS count FROM customers GROUP BY gender; Result: The dataset contains 3425 male customers and 3555 female customers.
- (Repeat for additional queries and insights.)

Interactive Map with Folium Results

- Geographical Distribution: Our interactive map displayed the geographical distribution of churned customers, highlighting clusters of high churn rates in urban areas.
- Heatmaps: Utilizing Folium's heatmap feature, we visualized the density of customer churn across different regions, identifying hotspots for further investigation.

Plotly Dash Dashboard Results

- Dashboard Overview: Our Plotly Dash dashboard provided interactive visualizations of key metrics such as churn rates by customer demographics, subscription plans, and service usage.
- Key Metrics: The dashboard also included metrics such as customer lifetime value (CLV) and net promoter score (NPS) to assess overall customer satisfaction and loyalty.

Predictive Analysis (Classification) Results

- Model Evaluation: Our logistic regression model achieved 85% accuracy in predicting customer churn, with a precision of 0.82 and a recall of 0.75.
- Confusion Matrix: The confusion matrix revealed 350 true positives, 120 false positives, 80 false negatives, and 1450 true negatives, indicating a relatively balanced performance of the model.
- Feature Importance: Analyzing feature importance identified monthly charges, contract type, and tenure as the most influential factors contributing to customer churn.

Conclusion

In conclusion, our project successfully analyzed customer churn patterns and developed a predictive model to identify at-risk customers. The insights gained from this analysis can help businesses implement targeted retention strategies and improve overall customer satisfaction.