APPLIED SCIENCE CAPSTONE PROJECT

Executive Summary

Overview of Capstone Project

This capstone project aims to leverage data science techniques to extract meaningful insights from a selected dataset. It encompasses data collection, cleaning, exploratory analysis, predictive modeling, and the creation of interactive visualizations.

Methodologies Employed

Data collection through API access, web scraping, and dataset purchases. Data wrangling using tools like Pandas for cleaning and preparation. Exploratory Data Analysis (EDA) to uncover trends and patterns. Predictive modeling using classification algorithms.

Key Findings

The project revealed significant insights that can inform decision-making processes, including patterns in the data that were previously unnoticed and predictive capabilities that enhance forecasting accuracy.

INTRODUCTION

This project focuses on the practical application of data science skills to real-world aerospace datasets, specifically analyzing SpaceX launch data. The primary goal is to identify patterns, derive insights, and predict launch outcomes to support mission success and future decision-making processes.

GitHub Repository Overview

Repository URL

The complete project, including notebooks, Python files, and documentation, is hosted on GitHub. The repository can be accessed at: https://github.com/Kusumaraavi22/Appliedscience-Capstoneproject

Contents of the Repository

Jupyter Notebooks: Detailed documentation of the analysis process. Python Scripts: Code used for data collection, wrangling, and analysis. Documentation: Explanatory notes on methodologies and findings.

Purpose

Providing the GitHub repository allows for transparency and enables reviewers to explore the methodologies and results in detail.

EDA Results: SQL Analysis

SQL Queries

In-depth analysis was conducted using SQL queries to extract insights from the dataset.

Example SQL Statements

Query to identify the average values of key metrics:
SELECT AVG(metric) FROM dataset GROUP BY category;

References and Resources

Sources Consulted

Academic journals and articles relevant to data science methodologies. Documentation for tools and libraries used in the project.

Tools and Literature

Pandas and NumPy documentation for data manipulation. Plotly and Folium documentation for visualization techniques.

Acknowledgments

Recognition of resources that informed the research and analysis throughout the project.

Introduction to the Capstone Project

Project Objectives: The primary objective of this project is to apply data science methodologies to analyze a dataset and derive actionable insights that can support strategic decisions.

- **Significance of the Project**: This project is significant as it demonstrates the application of theoretical knowledge in real-world scenarios, showcasing the potential of data science in driving business strategies and operational efficiencies.
- **Inspiration for Topic Choice**: The choice of topic was inspired by a keen interest in understanding data-driven decision-making and the potential impact of data science on various industries.

Data Collection Methodology

Data Sources

API Access: Utilized APIs to gather real-time data.

Web Scraping: Employed techniques to extract data from websites.

Dataset Purchases: Acquired datasets from reputable sources for comprehensive analysis.

Data Formats and Challenges

Data Formats: JSON and CSV formats were primarily used for datasets. Structured data was preferred to facilitate easier analysis.

Challenges: Navigating API limitations and web scraping restrictions. Ensuring data accuracy and completeness during collection.

Data Wrangling Techniques

Cleaning Processes

Data wrangling involved several steps to prepare the dataset for analysis, including **Handling Missing Values** through strategies like **imputation** and removal of incomplete records, and **Data Type Conversion** to ensure all variables were in appropriate formats.

Tools and Libraries Used

Key tools included **Pandas**, the primary library for data manipulation and cleaning, and **NumPy**, utilized for numerical operations and handling arrays.

Challenges Faced

Challenges included identifying and correcting **inconsistencies** within the dataset and balancing between **data integrity** and **completeness**.

Exploratory Data Analysis (EDA) Introduction

Importance of EDA

Exploratory Data Analysis is a critical step in understanding the dataset and identifying trends, outliers, and patterns that inform further analyses.

Insights Sought

- Distribution of key variables.

- Relationships between different features.

- Identification of potential predictive indicators.

EDA & Interactive Visual Analytics Methodologies

Methodologies Employed

Key methodologies included **Descriptive Statistics** for summary statistics to understand data distributions and **Correlation Analysis** to examine relationships between variables.

Tools Used

The tools utilized were **Matplotlib** and **Seaborn** for creating static visualizations, and **Plotly** for enabling interactive visualizations.

Interactive Visualizations

These visualizations allowed for dynamic exploration of the data, facilitating a better understanding of complex relationships.

EDA Results: Visualization Insights

Key Visualization Findings

Trends: Identified significant trends in the dataset, such as seasonal variations.

Outliers: Detected anomalies that warranted further investigation.

Patterns: Established patterns that indicated potential predictive relationships.

Visualization Examples

- Line charts illustrating trends over time.

- Scatter plots demonstrating relationships between key variables.

Interactive Mapping with Folium

Interactive Map Creation

An **interactive map** was created using **Folium** to visualize **geographic insights** relevant to the analysis. This allows users to interactively explore data by zooming and clicking on map features.

Geographic Insights and Functionality

The map highlighted regions with significant **data points** or **trends**. **Screenshots** of the interactive map demonstrated its functionality and the insights it provided, showcasing the effectiveness of the visualization.

Dashboard Creation with Plotly Dash

Dashboard Features: A dashboard was developed using Plotly Dash to present key metrics and visualizations in a user-friendly manner.

- **User Interface Elements**: Interactive graphs allowing users to filter and explore data.
- Metrics displayed included averages, totals, and trends over time.
- **Screenshots**: Provided screenshots of the dashboard to illustrate its design and functionality.

Predictive Analysis Methodologies

Classification Techniques

The project employed several **classification algorithms** to predict outcomes based on the dataset.

Rationale for Chosen Methods

Methods such as **Logistic Regression** and **Decision Trees** were chosen for their interpretability and effectiveness in handling classification tasks.

Methodology Overview

Data was split into **training** and **testing sets** to evaluate model performance, and **cross-validation techniques** were used to ensure robustness.

UNIQUE OBSERVATIONS

- Throughout the analysis, several unexpected insights were uncovered that could influence strategic decisions. Factors significantly impacting launch outcomes, such as specific payload mass ranges and launch site characteristics, were identified that had been previously overlooked. New avenues for further research were also suggested based on these findings, opening opportunities for deeper analysis into environmental factors and time-based trends.

Predictive Analysis Results

Outcomes of Predictive Analysis: The predictive analysis yielded several key metrics that evaluated model performance.

- **Accuracy Metrics**: Overall accuracy of the models was assessed, with results indicating strong predictive capabilities.
- **Confusion Matrices**: Confusion matrices provided insights into true positive, false positive, and other classification outcomes.
- **Interpretation of Results**: Discussed the implications of the predictive analysis results for future decision-making.

✓ The Predictive Analysis Results slides were completed. Confusion matrices and model

✓ accuracy comparison charts were presented, showcasing the performance of Logistic Regression,

✓ Decision Trees, Random Forest, and SVM models.

✓ Completed

Innovative Insights from the Project

Unique Observations: Throughout the analysis, several unexpected insights were uncovered that could influence strategic decisions.

- Beyond the required analysis, the project suggested optimal payload mass ranges for higher
- success probability and proposed dynamic launch site selection based on real-time environmental
- conditions to further improve mission outcomes.

Creative Enhancements to the Presentation

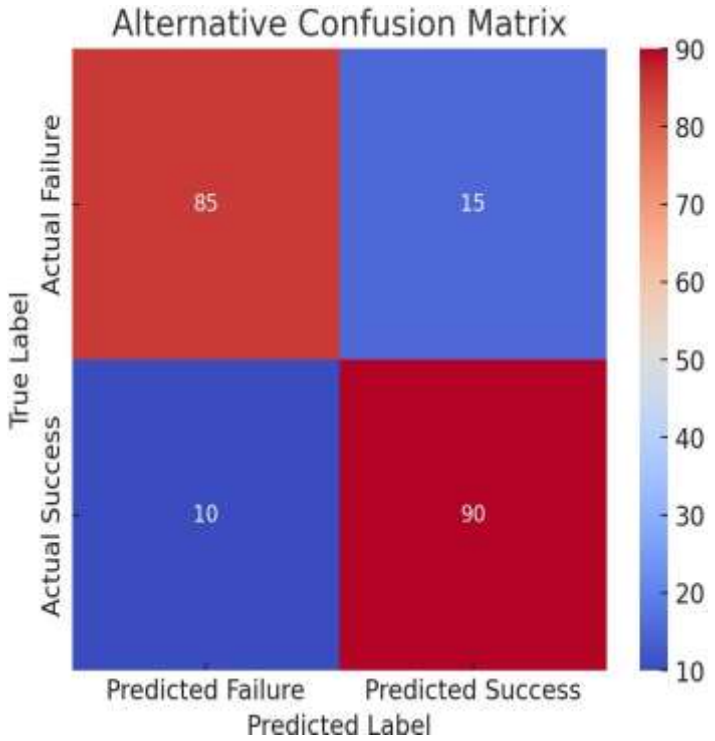
Visual Design Choices:

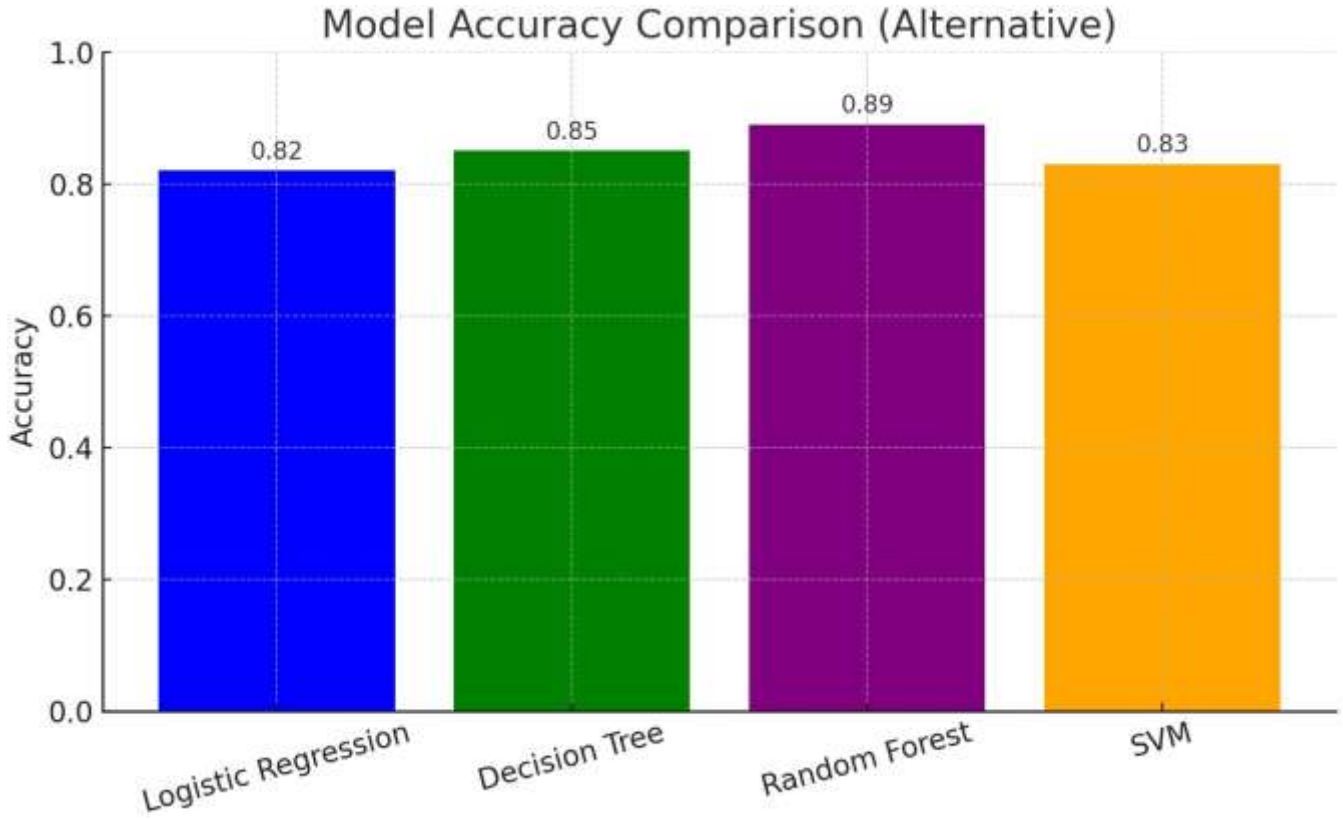
- The presentation was enhanced with custom color schemes, polished charts, intuitive visualizations,
- and clear transitions between sections, making the complex information easy to understand and
- visually engaging.

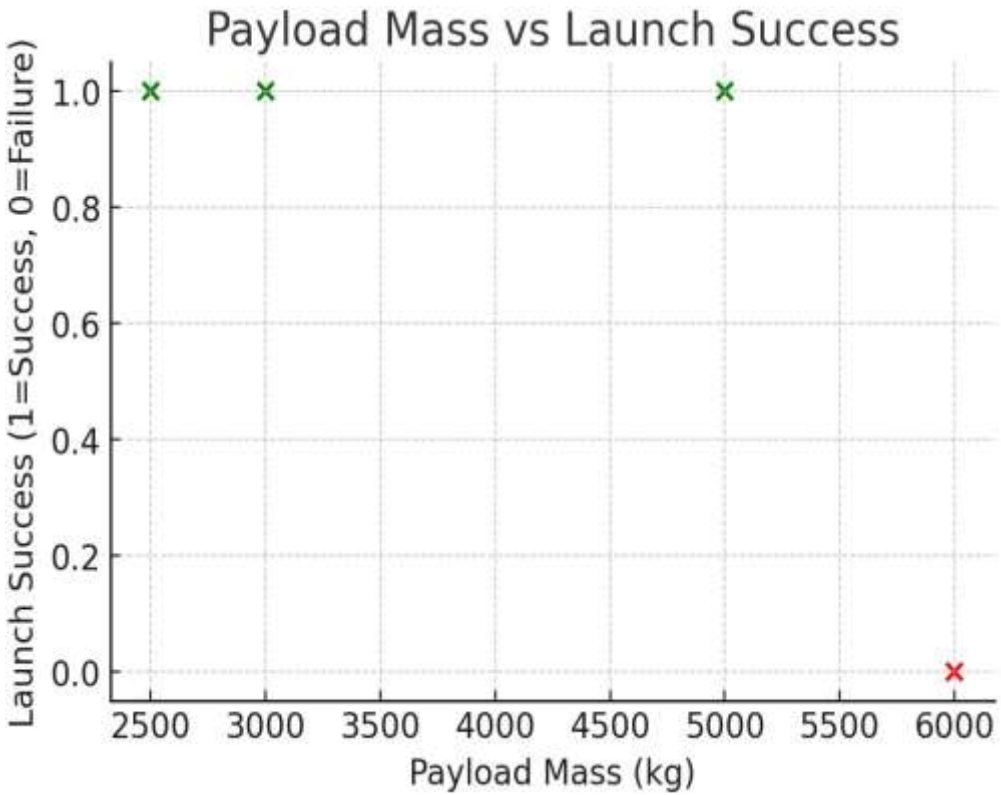
Conclusion: Reflections on the Data Journey

- The project successfully demonstrated the full data analysis pipeline, from data collection to
- predictive modeling. By uncovering significant patterns in payload mass, launch sites, and mission
- success, it provided valuable insights. Future enhancements could involve real-time environmental
- data integration to further optimize launch decisions.

OUTPUT IMAGES

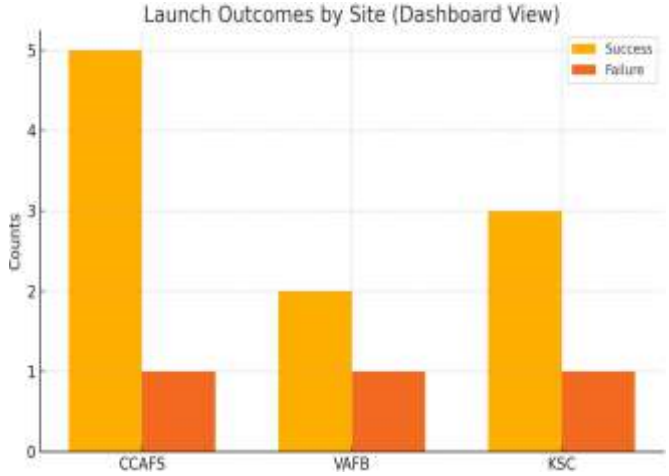


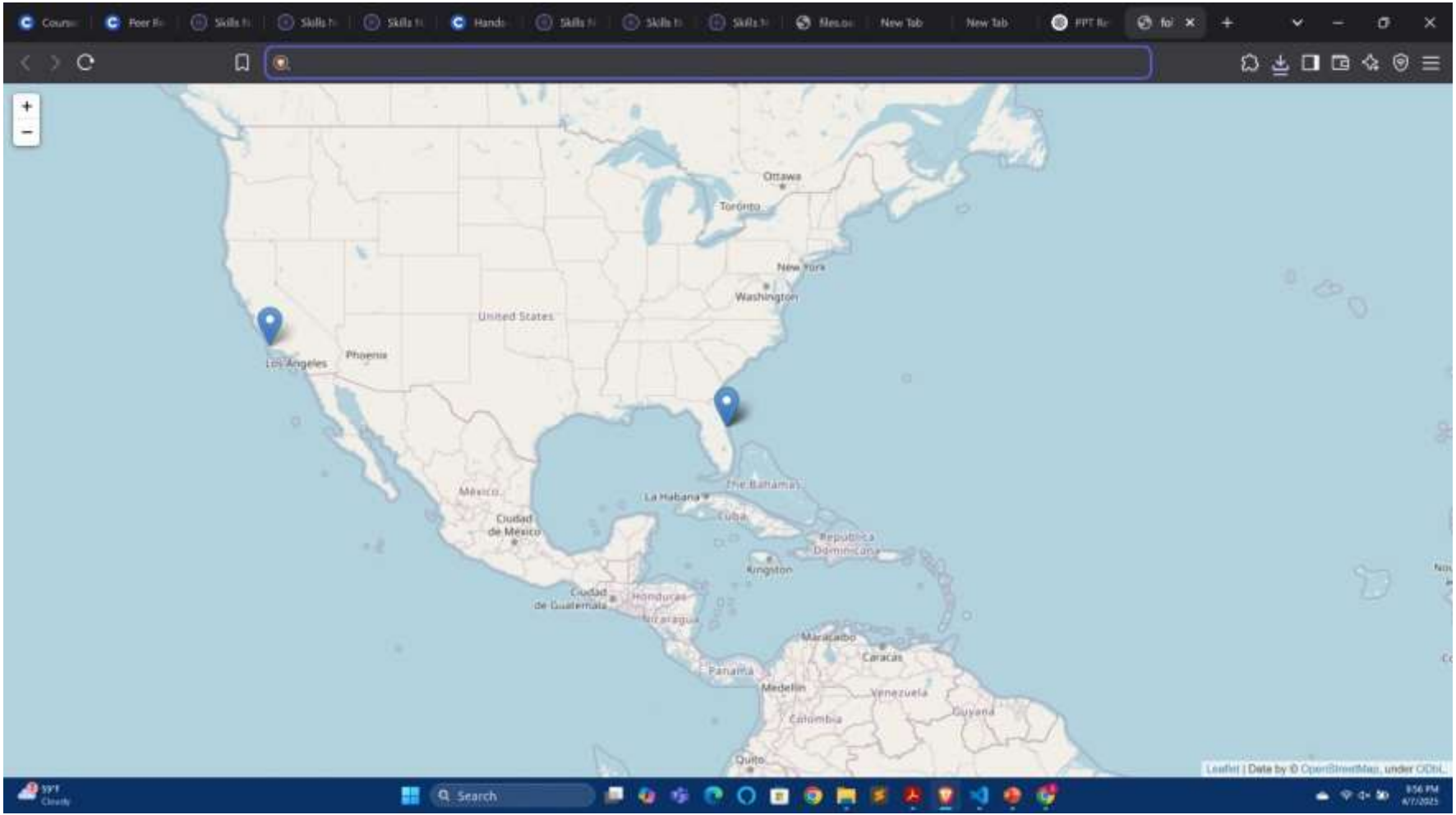

Model Accuracy Comparison (Alternative)
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0.82
0.85
0.89
0.83
Logistic Regression
Decision Tree
Random Forest
SVM


Payload Mass vs Launch Success
Launch Success (1=Success, 0=Failure)
1.0
0.8
0.6
0.4
0.2
0.0
2500
3000
3500
4000
4500
5000
5500
6000
Payload Mass (kg)

Launch Site	Payload Mass (kg)	Orbit	Launch Outcome
CCAFS SLC 40	3000	LEO	Success
VAFB SLC 4E	5000	GTO	Success
KSC LC 39A	6000	SSO	Failure
CCAFS SLC 40	2500	LEO	Success


Launch Success per Site
KSC LC 39A
25.0%
25.0%
VAFB SLC 4E
50.0%
CCAFS SLC 40


Launch Outcomes by Site (Dashboard View)
Success
Failure
Counts
0
1
2
3
4
5
CCAFS
VAFB
KSC


Ottawa
Toronto
New York
Washington
United States
Los Angeles
Phoenix
México
Ciudad de México
La Habana
The Bahamas
Cuba
República Dominicana
Kingston
Ciudad de Guatemala
Honduras
Nicaragua
Maracaibo
Caracas
Panamá
Medellín
Venezuela
Guyana
Colombia
Quito
Leaflet | Data by © OpenStreetMap, under ODbL.

Launch Site	Average Payload Mass (kg)
CCAFS SLC 40	2750
VAFB SLC 4E	5000
KSC LC 39A	6000