A COMPREHENSIVE JOURNEY FROM DATA COLLECTION TO ACTIONABLE INSIGHTS.

✷ The Data Science Capstone Project Luis Hdz explores data science methodologies applied to real-world challenges. It involves collecting, cleaning, analyzing, & visualizing data to derive insights. Documented on GitHub, this project applies theoretical knowledge in practical scenarios, leading to insights that drive strategic decisions.

MACHINE LEARNING MODELS ANALYSIS

DATA COLLECTION AND PREPARATION

Data from SpaceX API classified successful landings using SQL, visualization, folium maps, and dashboards.

MODEL DEVELOPMENT AND RESULTS

Developed four models with 83.33% accuracy; overpredicted successful landings.

COMMERCIAL SPACE AGE AND CHALLENGES

SPACEX'S COMPETITIVE EDGE

SpaceX offers launches for $62 million by recovering Stage 1 of rockets.

SPACE Y'S AMBITION

Space Y aims to rival SpaceX by predicting Stage 1 recovery success.

MACHINE LEARNING IN SPACE

Tasked to train a model for successful Stage 1 recovery prediction.

DATA COLLECTION AND ANALYSIS METHODOLOGY

DATA SOURCES

Combined data from SpaceX public API and SpaceX Wikipedia page.

DATA WRANGLING

Performed data wrangling to prepare the dataset for analysis.

CLASSIFICATION PROCESS

Classified true landings as successful and unsuccessful otherwise.

EXPLORATORY DATA ANALYSIS

Conducted EDA using visualization and SQL for insights.

INTERACTIVE VISUAL ANALYTICS

Utilized Folium and Plotly Dash for interactive visual analytics.

PREDICTIVE ANALYSIS

Performed predictive analysis using classification models, tuned with GridSearchCV.

Overview

DATA-DRIVEN DECISIONS REQUIRE ROBUST METHODOLOGIES.

Our approach focuses on a comprehensive methodology that includes data collection, wrangling, visualization, dashboard creation, & modeling. This ensures data is accurate, insightful, & actionable. We collect data from various sources, clean & prepare it, apply visualization techniques, integrate insights into dashboards, & develop predictive models to forecast trends.

DATA COLLECTION OVERVIEW

API DATA COLLECTION

Involves API requests from Space X public API to gather data such as FlightNumber, BoosterVersion, LaunchSite, and more.

WEB SCRAPING PROCESS

Web scraping performed on Space X's Wikipedia entry to collect data columns like Flight No., Payload, Customer, and Date.

NEXT STEPS: API FLOWCHART

The next slide will illustrate the flowchart detailing the data collection process from the Space X API.

UPCOMING: WEB SCRAPING FLOWCHART

The following slide will present the flowchart for the web scraping data collection methodology.

DATA PROCESSING WORKFLOW

DATA COLLECTION

Utilize SpaceX API to gather data on SpaceX launches.

Request SpaceX APIs
.JSON file
Lists of Launch Site, Booster Version, Payload Data

DATA TRANSFORMATION

Convert JSON file data into a structured DataFrame format.

Json_normalize to DataFrame
Dictionary of relevant data

DATA CONVERSION

Cast dictionary into a DataFrame and prepare for filtering.

DataFrame from dictionary

DATA FILTERING AND IMPUTATION

Filter data for Falcon 9 launches and handle missing values.

Filtered Falcon 9 data
Imputed PayloadMass values using mean

DATA COLLECTION: WEB SCRAPING

REQUEST HTML

Use a script to send a request to Wikipedia's server to obtain the HTML content of a web page.

HTML content of Wikipedia page

PARSE HTML

Utilize BeautifulSoup with html5lib parser to navigate and parse the HTML structure.

Parsed HTML document

EXTRACT DATA

Identify and extract the launch information from the HTML table, converting it to a dictionary.

Dictionary of launch data

CONVERT TO DATAFRAME

Transform the extracted dictionary data into a structured DataFrame for analysis.

Pandas DataFrame containing launch information

PREDICTIVE ANALYSIS WORKFLOW

- Begin with splitting the label column ‘Class’ from the dataset to separate features from the target variable.
- Apply Standard Scaler to fit and transform the features, ensuring data is standardized before model training.
- Implement GridSearchCV with cross-validation (cv=10) to determine the optimal parameters for each model.
- Evaluate models on the split test set to measure their predictive performance.
- Create a barplot to compare the scores of the models, providing a visual representation of their effectiveness.

Results

EXPLORING DATA ANALYSIS RESULTS

OVERVIEW OF ANALYSIS TECHNIQUES

- Preview of the Plotly dashboard showcasing interactive graphs and charts.
- Exploratory Data Analysis (EDA) using visualization techniques provides insights into data patterns and trends.
- EDA with SQL was employed to query and manipulate large datasets efficiently.
- An interactive map created with Folium highlights key geographical data points.
- Model results indicate an accuracy of approximately 83%, reflecting robust predictive capabilities.