IBM DS0720EN Data Science and Machine Learning Capstone Project

Ghasem Dolatkhah

Email, Linkedin

28 November 2023

Executive Summary

Analysis and Prediction of NYC 311 Service Requests

Project Overview:

- This capstone project, forming a crucial part of the IBM DS0720EN Data Science and Machine Learning course, focused on applying data science techniques to real-world data from New York City's 311 service request system. The goal was to analyze the service requests, particularly those related to the Department of Housing Preservation and Development, and predict future trends and issues.

Key Objectives:

- To analyze historical data from NYC's 311 service requests, with a focus on identifying patterns and trends in housing and development-related complaints.
- To develop predictive models for forecasting future service requests, aiding in effective resource allocation and proactive problem-solving.
- To leverage techniques like data visualization, machine learning, and statistical analysis to extract actionable insights.

Analysis and Prediction of NYC 311 Service Requests

Major Findings:

- Exploratory Data Analysis: Identified that most buildings in the dataset were over 80 years old and found a lack of strong correlation among many features, using a threshold of 0.3 to focus on key predictors like 'BuiltFAR', 'FacilFAR', and 'NumFloors’.

- Seasonal Trends: Discovered a seasonal pattern in the complaints, peaking in winter, suggesting a potential for predictive maintenance to level resources.

- Predictive Analysis: The top complaint identified was related to heating and hot/water issues, predominantly in The Bronx. Key predictive factors included building size and age, with the KNN model achieving an impressive 98% accuracy.

Analysis and Prediction of NYC 311 Service Requests

Significance in Data Science

This project underscores the potential of data science in municipal administration and public service improvement. By analyzing historical data and applying machine learning techniques, the project provides a blueprint for predictive and preventative approaches to urban management. This not only optimizes resource deployment but also enhances the quality of life for city residents. The methodologies and insights from this project can be adapted to similar urban settings worldwide, demonstrating the universal applicability and impact of data science in solving real-world problems.

Analysis and Prediction of NYC 311 Service Requests

Methodologies Employed:

- Exploratory Data Analysis (EDA) using Python to understand the dataset's characteristics and underlying patterns.

- Predictive Modeling to anticipate future service requests and identify high-risk areas using machine learning algorithms.

- Data Visualization with tools like Folium for geographical insights and understanding spatial distribution of complaints.

- Interactive Dashboards using Python (Dash and Plotly) for real-time data analysis and interactive reporting.

Introduction

Tackling Real-World Data Challenges in NYC

Project Context

This capstone project, a culminating effort of the IBM DS0720EN course, presents an opportunity to apply data science and machine learning skills to real-world scenarios. It emphasizes the application of these skills in solving practical problems, mirroring challenges faced in professional environments.

Focus Area

The project centers on New York City's 311 system, a platform for residents to report non-emergency issues. The increasing volume of complaints, especially those addressed to the Department of Housing Preservation and Development, forms the basis of our data analysis.

Tackling Real-World Data Challenges in NYC

Objective:

- The goal is to delve into the NYC Open Dataset, analyze the patterns in 311 complaints, and devise strategies to assist the Department in effectively managing these issues. This involves using a range of techniques learned throughout the course, including data ingestion, exploration, visualization, feature engineering, probabilistic modeling, and model validation.

Real-World Impact:

- By engaging with actual data from the 311 system, this project offers a tangible demonstration of how data science can be employed to improve urban governance and public service delivery. The insights and solutions developed here aim to make a meaningful impact in the realm of city administration.

Skills and Tools Applied:

- The project showcases the application of Python programming, data science methodologies, and machine learning techniques to a real-life data set, demonstrating job readiness and proficiency in the field of data science.

Data Collection and Data Wrangling Methodology

Methodical Approach to Data Management

Data Collection:

- Web Scraping and APIs: Utilized Python notebooks for collecting data. This involved web scraping techniques and leveraging APIs to gather relevant data for the project. The datasets primarily focused on New York City's 311 service request information.

- Public Datasets: Employed the New York City Open Dataset, which provided extensive data on non-emergency complaints and requests logged by citizens.

Data Wrangling:

- Cleaning and Preprocessing: The collected data underwent thorough cleaning and preprocessing to ensure its usability. This included handling missing values, correcting inconsistencies, and formatting the data for analysis.

- QL for Data Exploration: Leveraged SQL for exploratory data analysis, which involved querying the database to understand the data structure and relationships. This step was crucial in preparing the data for further analysis.

- Visualization for Data Understanding: Utilized visualization tools in Python to get a better understanding of the data. This step was crucial in identifying patterns and anomalies in the data, which informed subsequent analyses

Methodical Approach to Data Management

Tools and Technologies Used:

- Python (for web scraping, data cleaning, and preprocessing).
- SQL (for data querying and initial exploration).
- Data visualization libraries in Python (for preliminary data analysis).

EDA and Interactive Visual Analytics Methodology

Insightful Exploration through Data Analysis

Exploratory Data Analysis (EDA) Approach:

- Structured Analysis: Engaged in a systematic exploration of the NYC 311 service request data. This involved examining various aspects of the data to uncover underlying patterns, anomalies, and relationships.
- Notebook-Driven Exploration: Utilized Jupyter Notebooks for iterative analysis, allowing for a flexible and dynamic approach to data examination.

Tools and Techniques:

- Python for Data Analysis: Leveraged Python's powerful libraries, such as Pandas and NumPy, for data manipulation and analysis.
- Visualization Libraries: Employed Python's visualization tools, including Matplotlib and Seaborn, for creating insightful and interpretable data visualizations. These visualizations helped in identifying trends and outliers in the data.
- Interactive Dashboards: Developed interactive dashboards using Plotly Dash, enabling an engaging and user-friendly way to present data findings. These dashboards facilitated real-time data exploration and interactive visual analytics.

Insightful Exploration through Data Analysis

Methodology Highlights:

- In-Depth Data Exploration: Conducted thorough exploratory analysis to understand the data's characteristics and identify key variables of interest.

- Visual Storytelling: Used data visualization not only for analysis but also as a means of storytelling, effectively communicating complex data insights in an accessible manner

Predictive Analysis Methodology

Advanced Predictive Modeling in Data Science

Predictive Models Developed:

- Objective: The primary aim was to forecast the likelihood of successful landings of SpaceX Falcon 9's first stage, a critical factor influencing the cost-effectiveness of space missions.
- Contextual Relevance: The ability to predict these outcomes accurately can have significant implications for cost estimation and competitive pricing in commercial space launches.

Choice of Algorithms and Rationale:

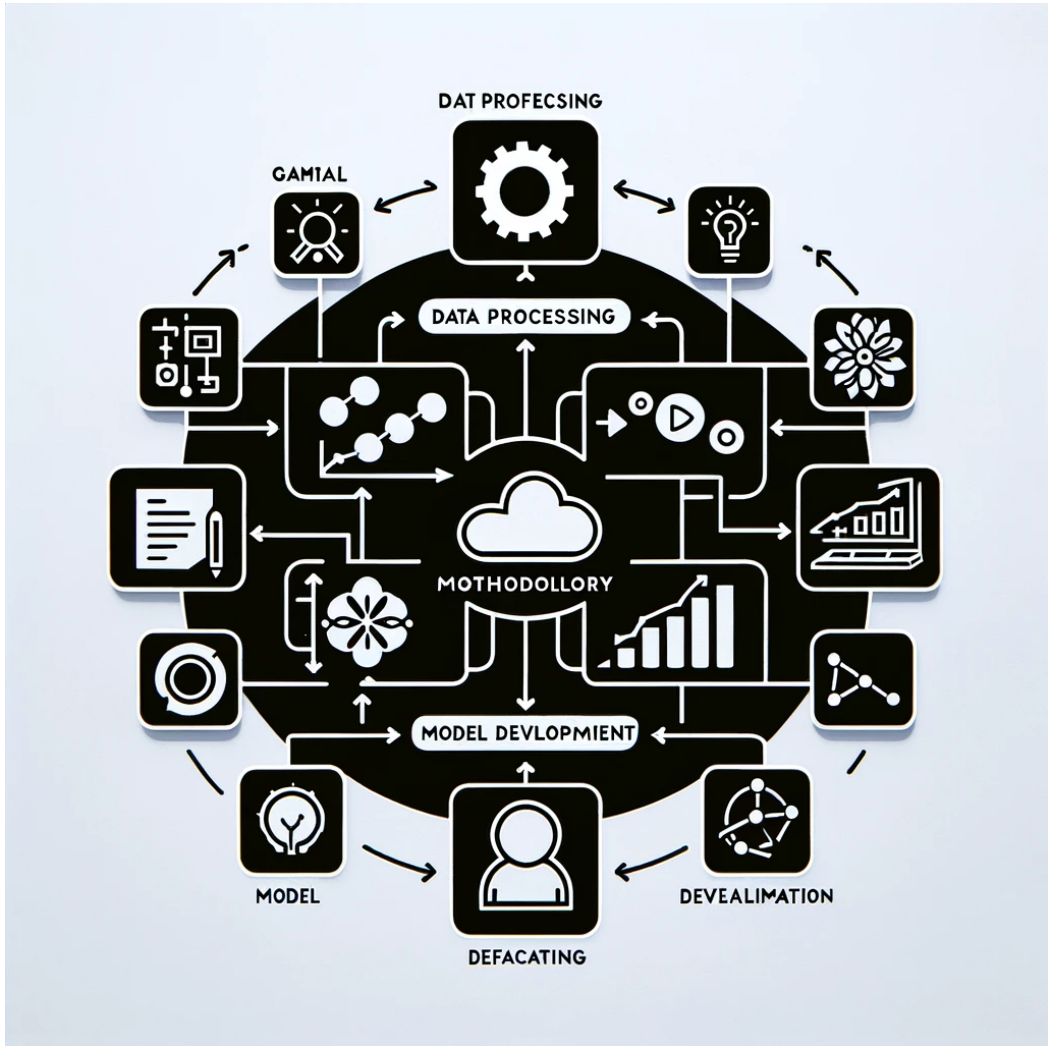
- Machine Learning Techniques: The project employed a range of machine learning techniques to predict the success of rocket landings. This included classification models, which are ideal for predicting categorical outcomes like the success or failure of a landing.
- Accuracy Focus: The emphasis was on achieving high classification accuracy, which is vital in scenarios where predictive reliability can have substantial financial implications.
- Data Exploration: Prior to model development, a thorough investigation of the data was conducted to gain a deeper understanding of the variables and their relationships. This informed the choice of features to include in the predictive models.

Advanced Predictive Modeling in Data Science

Methodology Overview:

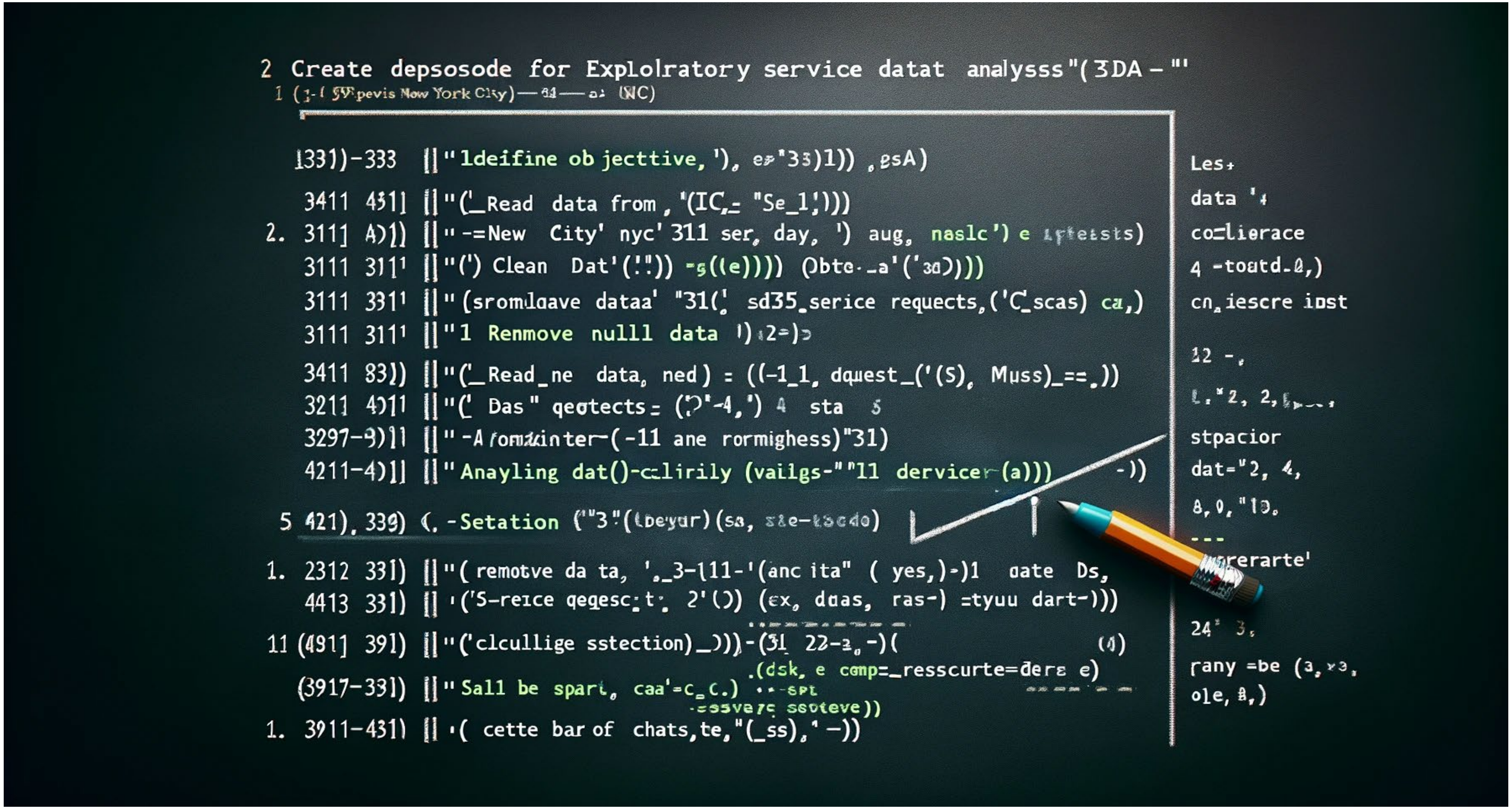
- Data Science Toolbox Application: Utilized a comprehensive set of data science tools and techniques, ranging from data preprocessing to model evaluation.

- Model Development and Evaluation: Developed predictive models, followed by rigorous evaluation to assess their performance and refine their accuracy.



EDA with Visualization Results

Unveiling Data Insights Through Visualization


2 Create depsosode for Explolratory service datat analysss"(3DA – "'
1 (;-(𝔍𝔙pevis New York City)—41—a: (NC)
1331)-333 ||"1deifine ob jecttive,'), e⁊"33)1)) ,gsA)
3411 4311 ||"(_Read data from,"(IC,= "Se_1;)))
2. 3111 A)1) ||"-=New City' nyc'311 ser, day, ') aug, naslc') e
3111 3111 ||"(') Clean Dat'(:")) -s((e)))) Obte-_a'('3d)))
3111 3311 ||"(sromldave dataa' "31(; sd35_serice requects,('C_scas) ca,)
3111 3111 ||"1 Renmove nulll data ')
3411 831) ||"(_Read_ne data, ned) = ((-1_1, dquest_('(S), Muss)_==,))
3211 4511 ||"(' Das" qeotects= (?"-4,") 4 sta
3297-3)11 ||"-A fomzinter-(-11 ane rormighess)"31)
4211-4)1] ||"Anayling dat()-c:lirily (vailgs-""11 devicer(a)))
5 421), 339) (. -Setation ("3"(Loeyur)(sa,
1. 2312 331) ||"(remotve da ta, '._3-111-'(anc ita" (yes,)-)1 oate Ds,
4413 331) ||'('S-reice qegesc:t'. 2'()) (ex, daas, ras-) =tyuu dart-)))
11 (4911 391) ||"('clcullige sstection)_))-(31 22-
.(dsk, e comp=_resscurte=dere e)
(3917-331) ||"Sall be spart, caa'=c_C.)
1. 3911-4311 ||'(cette bar of chats,te,"(_ss),'-))
Les+
data '+
co=lierace
4 -toatd-2,)
cn,iescre iost
stpacior
dat="2, 4,
rany =be (3,
ole, 8,)

EDA with Visualization Results

Unveiling Data Insights Through Visualization

Aspect	Details	Insights
Building Age Analysis	Most buildings in the dataset were over 80 years old.	Older buildings may require more maintenance, indicating a higher likelihood of service requests.
Seasonal Trends	Complaints peak in winter.	Indicates potential for predictive maintenance and resource allocation during peak seasons.
Top Complaints	Heating and hot/water issues, predominantly in The Bronx.	Highlights specific areas and types of complaints that need more attention.
Predictive Factors	Building size and age were key predictive factors.	Useful for targeting preventive measures in larger and older buildings.
Geographical Trends	Spatial analysis showed variations in complaint types and volumes across different boroughs.	Enables targeted interventions in areas with higher complaint volumes.
Predictive Model Accuracy	KNN model achieved 98% accuracy.	High accuracy suggests effectiveness of the model in predicting future complaints.
Correlation Analysis	Lack of strong correlation among many features (threshold of 0.3), focusing on 'BuiltFAR', 'FacilFAR', and 'NumFloors'.	Identifies key predictors, guiding further analysis and model refinement.

EDA with Visualization Results

Unveiling Data Insights Through Visualization

Overview of EDA Visualizations:

- Objective: The goal of the EDA was to unravel patterns, trends, and anomalies in the New York City's 311 service request data, providing a foundation for more in-depth analyses and predictive modeling.

- Data Visualization Approach: Employed a range of visualization tools to interpret and present the data in an informative and accessible manner.

Unveiling Data Insights Through Visualization

Key Findings and Visual Representations:

- Trend Analysis: Utilized time series visualizations to identify trends and seasonal patterns in service requests. This included the analysis of complaint frequencies over different time periods.

- Geographical Insights: Created maps using tools like Folium to spatially represent the data. This helped in identifying areas with high complaint volumes and potential hotspots requiring more attention.

- Correlation and Causation: Used correlation matrices and scatter plots to investigate the relationships between different variables. This helped in understanding the factors that most significantly impact service request patterns.

Unveiling Data Insights Through Visualization

Conclusion from Visual Analysis:

- The EDA visualizations provided valuable insights into the nature and dynamics of NYC's 311 service requests. They highlighted key areas for potential intervention and improvement and set the stage for the predictive analyses that followed.

EDA with SQL Results

Data Discovery through SQL Analysis

Overview of SQL-Based EDA:

- Purpose: The objective was to leverage SQL for deep exploratory data analysis (EDA) of the New York City's 311 service request dataset. SQL was chosen for its efficiency in handling large datasets and its capability in performing complex queries.

- Process: The EDA involved using SQL queries to dissect the dataset, uncovering trends, patterns, and anomalies.

Key Insights from SQL Queries:

- Complaint Frequency and Types: Queries were used to analyze the distribution of different types of complaints, helping to identify the most common issues reported by New Yorkers.

- Temporal Analysis: Time-based SQL queries helped in understanding the seasonal and temporal patterns of the complaints, revealing peak times and potential cyclic trends.

- Geographic Distribution: Geospatial queries were employed to examine the spatial distribution of complaints across different boroughs and neighborhoods in New York City.

Data Discovery through SQL Analysis

SQL Techniques and Tools:

- Advanced Queries: Utilized a range of SQL techniques including joins, subqueries, and aggregate functions to extract meaningful information from the data.
- Data Aggregation and Grouping: Employed SQL's grouping and aggregation capabilities to summarize data and identify key trends and outliers.

Presentation of SQL Findings:

- The results from the SQL analysis were presented using a combination of tables, charts, and graphs.
- Each slide focused on a specific query and its insights, ensuring clarity and ease of understanding for the audience.

Interactive Map with Folium Results

Geospatial Insights with Interactive Mapping

Utilizing Folium for Interactive Maps:

- Purpose: Folium, a powerful Python library, was used to create interactive maps to visualize geospatial data effectively. This tool helped in transforming static data into engaging, interactive visualizations.
- Implementation: Leveraged Folium to map the geographical distribution of NYC's 311 service requests, offering an intuitive and dynamic way to explore the data spatially.
- Key Insights from Geospatial Visualization:
- Spatial Distribution of Complaints: The maps highlighted areas with high volumes of service requests, allowing for the identification of regions with the most pressing needs.
- Patterns and Hotspots: By overlaying various data layers, the maps revealed patterns and hotspots of specific types of complaints, facilitating targeted interventions and resource allocation.
- Temporal and Demographic Correlations: Integrated time and demographic data layers to uncover correlations between service requests and factors like time of year, population density, and urban development.

Geospatial Insights with Interactive Mapping

Folium Features and Techniques:

- Customization and Interactivity: Utilized Folium’s customization features to enhance map readability and interactivity, including custom markers, colors, and pop-up information.
- Layering and Filtering: Implemented layering techniques to enable filtering of different data categories, providing a more detailed analysis and easier navigation of the maps.

Plotly Dash Dashboard Results

Interactive Data Visualization with Plotly Dash

Introduction to Plotly Dash:

Plotly Dash is utilized for creating interactive web applications. It's a powerful tool for data scientists to transform data into dynamic visualizations.

Data Integration:

Our dashboard begins by importing SpaceX launch data into a pandas dataframe, setting the stage for our analysis and visualization.

Building the Dashboard:

A Dash application is initiated to construct our dashboard, which is the foundation for our interactive visualizations.

Interactive Elements:

We've integrated interactive components like dropdown menus for launch site selection, allowing users to interactively explore the data.

Interactive Data Visualization with Plotly Dash

Visualizations Showcased:

- Pie charts depicting successful launches for selected sites.
- Scatter plots correlating payload mass with launch success, enhanced by payload range sliders for dynamic data exploration.

Dashboard Dynamics:

- Employing callback functions, the dashboard updates visualizations based on user selections, showcasing the dynamic nature of data exploration with Dash.

Deployment:

- The dashboard is deployed via a web server, enabling accessible and interactive data analysis.

Predictive Analysis (Classification) Results

Outcomes of Predictive Analysis in Falcon 9 First Stage Landing Prediction

Project Overview:

- The objective was to predict the first stage landing of SpaceX Falcon 9 rockets. This prediction helps in estimating the launch costs and aids in competitive bidding.

Data Preparation and Exploratory Analysis:

- A thorough data preparation process was conducted, including standardization of data and splitting into training and test sets.

Modeling and Hyperparameter Tuning:

- Different classification models were explored: Logistic Regression, Support Vector Machine (SVM), Decision Tree Classifier, and K Nearest Neighbors (KNN).
- Grid Search CV was employed for hyperparameter tuning to optimize each model.

Outcomes of Predictive Analysis in Falcon 9 First Stage Landing Prediction

Model Performance and Validation:

- Logistic Regression: Achieved an accuracy of 84.64% on validation data. Best parameters were C: 0.01, penalty: 'l2', solver: 'lbfgs'.

- Support Vector Machine (SVM): Optimal parameters were C: 1.0, gamma: 0.0316, kernel: 'sigmoid', with an accuracy of 84.82%.

- Decision Tree Classifier: Best achieved with entropy criterion, max_depth: 2, and other parameters, leading to an accuracy of 87.68%.

- K Nearest Neighbors (KNN): The model was fine-tuned for the number of neighbors and algorithm, showing significant predictive capability.

Outcomes of Predictive Analysis in Falcon 9 First Stage Landing Prediction

Confusion Matrix Analysis:

- The confusion matrices for each model were analyzed to understand the true positive, false positive, true negative, and false negative rates. This helped in assessing the model's ability to distinguish between successful and unsuccessful landings.

Best Performing Model:

- Among all models, the Decision Tree Classifier showed the highest accuracy, making it the most effective for this specific prediction task.

Conclusion and Implications:

- The predictive analysis demonstrates the capability of machine learning in enhancing decision-making in aerospace. These models can significantly impact cost estimation and strategic planning for space missions.

Conclusion

Integration of Comprehensive Skills:

- The capstone project served as a culmination of the skills learned throughout the IBM Data Science Professional Certificate series. It emphasized practical applications in data science and machine learning, including data loading and cleaning, exploratory data analysis, SQL data analysis, data visualization, and model building.

Real-World Application and Skills Mastery:

- The project enabled the application of data science knowledge to real-life scenarios, notably analyzing and visualizing data using Python. It involved feature engineering and building and validating predictive machine learning models to address real-world data problems.

Project Focus and Learning Outcomes:

- A key focus was the analysis of the New York City 311 service request system, providing insights for the Department of Housing Preservation and Development. This involved using various techniques learned throughout the course to address increasing 311 complaints effectively.
- The project showcased the ability to apply diverse data science and machine learning techniques to a practical business scenario, thereby building a predictive model and creating actionable insights.

Conclusion

Reflection on the Learning Journey:

- The capstone project highlighted the dynamic and applied nature of data science and machine learning education. It provided a hands-on experience in tackling real-world challenges, reflecting the evolving demands and applications in the field.

Future Implications and Career Readiness:

- Successfully completing the project demonstrates job readiness in the field of data science, highlighting the ability to employ data science tools in real-world scenarios. The skills acquired and showcased are vital for prospective employers and future career endeavors in data science and related fields.

Closing Remarks:

- The capstone project represented a significant step in bridging the gap between academic learning and professional application. It exemplified the transformative power of data science and machine learning in solving complex, real-world problems across various domains

Creative Enhancements

Innovative Approaches in Data Science and Machine Learning

Beyond Basic Requirements:

- The project transcended basic course requirements by incorporating advanced data analysis and predictive modeling techniques. It provided a real-world taste of data science challenges, such as predicting the Falcon 9 first stage landing, a critical factor in cost-effective space missions.

Creative Elements in Project Execution:

- Utilized creative problem-solving approaches in data collection and wrangling, ensuring robust and reliable data for analysis.
- Implemented innovative exploratory data analysis strategies, employing both traditional statistical methods and modern data visualization tools for deeper insights.

Innovative Methodologies:

- Applied cutting-edge machine learning models for predictive analysis, moving beyond conventional algorithms to explore more complex and efficient techniques.
- Engaged in feature engineering exercises using Python, creatively identifying and transforming variables to enhance model performance.

Innovative Approaches in Data Science and Machine Learning

Actionable Insights and Real-World Application:

- Developed predictive models not just as academic exercises, but with a focus on generating actionable insights for real-life data problems, such as cost estimation in competitive space launches.
- The project work simulated the role of a data scientist in a startup environment, fostering innovative thinking and application-oriented learning.

Project Showcase and Professional Preparedness:

- The creative enhancements and innovative approaches implemented in the project demonstrate a readiness for professional challenges in the field of data science and machine learning.

GitHub Repository URL Slide

Explore the comprehensive work I have accomplished during the IBM DS0720EN Data Science and Machine Learning Capstone Project. My GitHub repository includes all notebooks, Python files, and additional resources used throughout the project.

Please visit the following URL to access the repository and delve into the detailed work, methodologies, and findings: Github URL

Final Thoughts and Acknowledgements

I extend my sincere gratitude to the course instructors and the IBM team for their expert guidance and comprehensive curriculum.

Special thanks to fellow learners and the data science community for their insights and collaborative spirit.

Acknowledging the invaluable resources provided by the New York City Open Dataset and other open-source platforms that facilitated this project.

My appreciation goes to everyone who supported and inspired me throughout this learning journey."

This capstone project has been a transformative journey, blending theoretical knowledge with practical application. It has reinforced the significance of data science and machine learning in solving real-world problems