Leveraging Predictive Analytics for Enhancing Reusability and Cost-Efficiency in Space Launch Operations: A Case Study of SpaceX's Falcon 9

Abhay Singh

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Zhen Zhu, for his invaluable guidance, support, and encouragement throughout the duration of this research. His expertise and insights have been instrumental in shaping the direction and depth of this dissertation. Professor Zhu's willingness to challenge my ideas while providing thoughtful feedback has not only enhanced the quality of this work but has also greatly contributed to my growth as a researcher.

I am also thankful to my family and friends for their unwavering support and understanding during this demanding journey. Their encouragement has been a constant source of motivation for me.

Table of Contents

List of Figures, Visualisations and Tables

List of Figures and Visualisations:

List of Tables:

Abstract

The purpose of this research study is to investigate the factors influencing the success rate of Falcon 9 rocket landings and to explore how predictive analytics can enhance cost-efficiency strategies for space launch companies. With the rapid advancements in space technology and the increasing importance of cost-effective operations, understanding these factors is crucial for improving launch success rates and reusability *(Foust, 2017)*.

The research employs a comprehensive methodology encompassing data collection, exploratory data analysis, and the application of various machine learning models. Data was sourced from Kaggle *(Sagar Varandekar, 2022)*, containing SpaceX's historical launch records, and key features such as payload mass, grid fins, legs, reused count, block, and launch site were examined. Models including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Neural Network, XGBoost, and K-Nearest Neighbors were trained and evaluated to predict the success of Falcon 9 landings.

The findings reveal that the Random Forest model outperforms others, achieving perfect accuracy. Significant predictors of successful landings include the presence of grid fins, the number of times a booster has been reused, and specific launch sites. Moreover, the research highlights that successful landings are strongly correlated with higher reusability, leading to substantial cost savings. The XGBoost model also demonstrated high accuracy, reinforcing the robustness of ensemble methods in predictive analytics.

The implications of this research are significant for space launch companies. By employing predictive analytics, companies can optimise their launch planning and recovery strategies, thus enhancing cost-efficiency and sustainability. The study contributes to the Sustainable Development Goals (SDGs) by promoting innovation in space technology (SDG 9), supporting climate action through efficient resource use (SDG 13), and fostering partnerships for technological advancements (SDG 17) *(United Nations, 2015)*.

This research study provides a comprehensive analysis of the technical and operational factors influencing Falcon 9 rocket landings and demonstrates the potential of predictive analytics in revolutionising space launch operations.

1. Introduction

Space exploration has witnessed remarkable advancements in recent years, driven by both governmental space agencies and private companies. Among these private entities, SpaceX has emerged as a leader in innovation, particularly with its Falcon 9 rocket *(Wikipedia Contributors, 2024)*. This research study aims to investigate the factors influencing the success rate of Falcon 9 rocket landings and explore how predictive analytics can enhance cost-efficiency strategies for space launch companies.

The primary objective of this dissertation is to identify and analyse the technical and operational factors that contribute to the successful landing of Falcon 9 rockets. Additionally, the research aims to demonstrate how predictive analytics can be utilised to enhance cost-efficiency in space launch operations. By leveraging historical launch data, the study seeks to provide actionable insights that can improve the success rate of rocket landings and optimise reusability, ultimately leading to significant cost savings.

The successful landing and reusability of rockets are critical components in reducing the cost of space missions. However, achieving consistent success in these landings is fraught with challenges. Factors such as payload mass, launch site conditions, and specific technical features can significantly impact the outcome of a landing. Furthermore, the high costs associated with space launches necessitate a thorough understanding of these factors to develop effective cost-efficiency strategies *(Wikipedia Contributors, 2024)*. This research addresses the need for a comprehensive analysis of these variables and the application of predictive analytics to enhance decision-making processes in space operations.

The research methodology adopted in this study includes data collection, exploratory data analysis, and the application of various machine learning models. Data was sourced from Kaggle *(Sagar Varandekar, 2022)*, encompassing key features such as payload mass, grid fins, legs, reused count, block, and launch site.

1.1 Research Methodology:

1. **Data Preparation:** Initial exploration of the dataset to handle missing values and feature engineering to create relevant variables, such as the SuccessfulLanding indicator.

2. **Model Training and Evaluation:** Implementation of multiple machine learning models, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Neural Network, XGBoost, and K-Nearest Neighbors. The models were trained and evaluated based on accuracy, precision, recall, and F1-score.

3. **Predictive Analytics:** Development of a predictive model to estimate cost savings based on successful landings and reusability, using a hypothetical cost savings variable.

4. **Scenario Analysis:** Assessment of different scenarios to understand potential cost savings under varying conditions.

1.2 Dissertation Structure:

- **Chapter 2 - Literature Review:** Provides an in-depth review of existing academic literature related to rocket landings, reusability, and predictive analytics in space operations.
- **Chapter 3 - Methodology:** Details the data collection process, feature engineering, and the methods used for analysis, including the machine learning models employed.
- **Chapter 4 - Findings and Analysis:** Presents the results of the data analysis and model evaluations, highlighting the key factors influencing rocket landings and the performance of the predictive models.
- **Chapter 5 - Discussion:** Discusses the implications of the findings, potential biases and limitations of the study, and how the results contribute to Sustainable Development Goals (SDGs).
- **Chapter 6 - Conclusion:** Summarises the main conclusions of the research, provides managerial recommendations, and suggests areas for future research.

2. Literature Review

2.1 Overview of Predictive Modelling in Aerospace and the Impact of SpaceX's Falcon 9

Predictive modelling in aerospace represents a pivotal innovation, increasingly recognised for its role in advancing the safety, efficiency, and cost-effectiveness of space missions. As space exploration grows more complex, the integration of advanced data science techniques becomes essential for optimising aerospace operations. This breakthrough will not only reduce the cost of space missions but also demonstrate the transformative potential of predictive analytics in aerospace engineering *(Brown, 2023; www.thespacereview.com, n.d.).*

This literature review examines several critical aspects of predictive modelling in aerospace. It begins by discussing the significance of predictive modelling, particularly in improving the reusability of rocket components and its impact on operational efficiency. The review then explores the methodologies employed in predictive modelling, focusing on key techniques such as data preprocessing and the application of machine learning algorithms. Following this, the role of predictive analytics in cost reduction strategies within the aerospace sector is analysed, showcasing its effectiveness in streamlining operations and reducing expenditures. Finally, the review discusses the competitive implications of predictive modelling, providing insights into how this technology offers strategic advantages in a highly competitive industry.

2.2 The Impact of Predictive Analytics on Rocket Reusability and Shaping Aerospace Innovation

Predictive modelling has become an essential tool in the aerospace sector, particularly in the context of enhancing the reusability of rocket components. The increasing demand for cost-effective and sustainable space missions has propelled the adoption of predictive analytics, which enables the accurate forecasting of component wear and tear, mission success rates, and other critical variables *(Aditya Singh Tharran, 2023).* This technology allows aerospace companies to not only anticipate and mitigate potential risks but also optimise the performance of their systems *(Arnold, 2023).*

One of the most significant applications of predictive modelling in aerospace is its role in the reusability of rocket stages. The Falcon 9 rocket, developed by SpaceX, is a prime example of how predictive analytics can revolutionise the industry. By accurately predicting the conditions under which the rocket's first stage can be safely recovered and reused, SpaceX has dramatically reduced the cost of space missions *(Review, 2024).* This achievement not only demonstrates the practical benefits of predictive modelling but also sets a new benchmark for the industry. The successful reusability of the Falcon 9 has proven that rockets can be flown multiple times, significantly lowering the cost per launch and making space more accessible *(SpaceX, 2024; Jo and Ahn, 2021; Boiani, 2021).*

Moreover, predictive modelling contributes to operational efficiency by enabling more informed decision-making. For instance, the use of predictive analytics in mission planning allows for the optimisation of fuel usage, trajectory planning, and payload management (*Tománek and Hospodka, 2018).* This leads to more efficient missions with higher success rates, further enhancing the sustainability of space exploration efforts. Additionally, predictive models can help in the early detection of potential failures in aerospace components, allowing for proactive maintenance and reducing the likelihood of costly and catastrophic mission failures *(Prous, n.d.).*

The impact of these advancements extends beyond cost savings. By improving the reliability and safety of space missions, predictive modelling also plays a crucial role in advancing the overall goals of space exploration. The ability to accurately forecast mission outcomes and component performance not only supports current space endeavours but also paves the way for more ambitious projects, such as manned missions to Mars and beyond *(Lionnet, 2021).* As such, the significance of predictive modelling in aerospace is evident in its capacity to enhance both the efficiency and ambition of space missions, setting new standards for the industry *(Baiocco, 2021).*

2.3 Exploring Advanced Techniques in Predictive Analytics for Aerospace Engineering

The methodologies employed in predictive modelling within aerospace engineering are diverse, encompassing a range of techniques from data collection and preprocessing to the application of sophisticated machine learning algorithms. These methodologies are integral to the development of predictive models that can accurately forecast outcomes and optimise aerospace operations.

2.3.1 Data Collection and Preprocessing

The first step in predictive modelling involves the collection and preprocessing of data, which is crucial for ensuring the accuracy and reliability of the models. In aerospace engineering, data collection often includes a wide array of variables such as rocket specifications, environmental conditions, mission parameters, and historical performance data. This data must be meticulously cleaned and pre-processed to eliminate noise and inconsistencies, which could otherwise lead to inaccurate predictions *(Kaur, 2023).*

Data preprocessing typically involves several steps, including data cleaning, normalisation, and feature selection. Data cleaning addresses any missing or incorrect data points, ensuring the dataset's integrity. Normalisation standardises the data, bringing all variables to a similar scale, which is essential for the effective functioning of many machine learning algorithms. Feature selection is another critical aspect, where relevant features are identified and selected based on their importance in predicting the desired outcome. In aerospace, features might include variables like thrust levels, payload weight, fuel capacity, and atmospheric conditions at the time of launch *(Sorini et al., n.d.; Sirohhi, A, 2024).*

2.3.2 Machine Learning Algorithms

Once the data is pre-processed, the next step involves the application of machine learning algorithms. These algorithms are used to analyse the data and generate predictive models that can forecast various outcomes in aerospace missions. Several machine learning techniques are commonly used in predictive modelling within aerospace, each with its specific applications and advantages.

- **Regression Analysis:** Regression techniques are widely used in predictive modelling to predict continuous outcomes. In aerospace, regression models can predict variables such as fuel consumption, flight duration, or the likelihood of a successful landing. Linear regression, for instance, is often employed for its simplicity and interpretability, while more complex models like polynomial regression can capture non-linear relationships between variables *(Dataheadhunters.com, 2024; Verma, 2024).*

- **Classification Algorithms:** Classification algorithms are used when the prediction involves categorical outcomes. For example, logistic regression can be used to classify the success or failure of a mission based on historical data. Decision trees and random forests are also popular in aerospace for their ability to handle complex, non-linear relationships and provide interpretable results *(Alfarhood et al., 2024).*
- **Anomaly Detection:** Anomaly detection algorithms are crucial in identifying unusual patterns or outliers in the data, which might indicate potential issues in aerospace systems. These algorithms can detect anomalies in real-time, allowing for immediate corrective actions, which is particularly important in mission-critical operations *(Stanton et al., 2022).*
- **Neural Networks:** For more complex predictive tasks, neural networks, particularly deep learning models, are employed. These models are capable of processing large volumes of data and identifying intricate patterns that simpler models might miss. In aerospace, neural networks can be used for tasks such as image recognition in satellite imagery or optimising flight paths in real-time *(Pang et al., 2023; Le Clainche et al., 2023).*

2.3.3 Specific Tasks and Algorithms in Aerospace

In aerospace engineering, certain predictive tasks are particularly important, and specific algorithms have been developed to address these tasks effectively.

- **Trajectory Prediction:** Predicting the trajectory of rockets and spacecraft is a critical task in aerospace. Algorithms such as Kalman filters and particle filters are often used to estimate and predict the trajectory of moving objects based on noisy sensor data. These algorithms are essential in ensuring that spacecraft follow their intended paths and reach their destinations safely *(Pang et al., 2023).*
- **Reliability Analysis:** Ensuring the reliability of aerospace components is vital for mission success. Reliability analysis often involves the use of survival analysis techniques, which predict the lifespan of components based on historical failure data. This approach helps in planning maintenance schedules and reducing the risk of component failure during critical missions *(Stanton et al., 2022).*
- **Optimisation Algorithms:** Optimisation is another key area in aerospace, where algorithms are used to optimise various aspects of mission planning and execution. Genetic algorithms, for example, are employed to optimise flight paths, fuel consumption, and payload distribution. These algorithms simulate the process of natural selection to find the most efficient solutions to complex optimisation problems *(Sirohhi, A, 2024).*

2.4 Leveraging Predictive Analytics for Cost Efficiency and Sustainability in Aerospace

Predictive analytics has emerged as a powerful tool in the aerospace sector, playing a crucial role in cost reduction strategies across various operational domains. The ability to anticipate potential issues and optimise processes through data-driven insights has enabled aerospace companies to achieve significant cost savings while maintaining high levels of operational efficiency and sustainability.

2.4.1 Predictive Maintenance

One of the most impactful applications of predictive analytics in cost reduction is predictive maintenance. By analysing data from aircraft systems, predictive models can forecast when a component is likely to fail, allowing for maintenance to be scheduled before the failure occurs. This approach not only prevents costly unplanned downtime but also extends the lifespan of components by ensuring they are maintained in optimal condition *(Muilwijk, 2024).*

The use of predictive maintenance is particularly evident in engine health monitoring (EHM), where AI-driven models analyse data from engines to detect early signs of wear or malfunction. This proactive approach to maintenance has been shown to significantly reduce maintenance costs and improve the reliability of aircraft. Also, airlines that have adopted predictive maintenance strategies have seen substantial cost savings by avoiding unnecessary maintenance and minimising aircraft downtime *(Dupont, 2024; Korba et al., 2023).*

2.4.2 Manufacturing Strategies

Predictive analytics also plays a vital role in aerospace manufacturing, where it is used to optimise production processes and reduce costs. By analysing data from manufacturing operations, predictive models can identify inefficiencies and suggest improvements that lead to cost savings. For example, predictive models can optimise the use of materials, reduce waste, and streamline production schedules, all of which contribute to lower manufacturing costs *(Kaplan, 2024).*

In addition to optimising existing processes, predictive analytics can also inform the design of new products. By simulating how different designs will perform under various conditions, predictive models can help engineers select the most cost-effective designs that meet performance and safety requirements. This approach not only reduces the cost of designing and developing new aerospace products but also ensures that these products are more reliable and easier to manufacture (*Anon, 2024; Teubert, Pohya and Gorospe, 2023).*

2.4.3 Operational Planning

Beyond maintenance and manufacturing, predictive analytics is also used to optimise operational planning in the aerospace sector. One area where this is particularly important is flight scheduling and routing, where predictive models can analyse various factors such as weather conditions, aircraft weight, and air traffic to recommend the most efficient flight paths. This optimisation reduces fuel consumption, shortens flight times, and lowers operational costs *(Kaplan, 2024).*

Furthermore, predictive analytics is used to forecast demand for parts and manage inventory levels more effectively, reducing excess inventory costs and ensuring that the necessary parts are available when needed (*Muilwijk, 2024).* In essence, predictive analytics is a powerful tool that drives cost efficiency across multiple facets of aerospace operations. By enabling more informed decision-making and optimising processes, predictive analytics not only helps aerospace companies reduce costs but also supports broader sustainability goals, ensuring the long-term viability of the industry *(Teubert, Pohya and Gorospe, 2023).*

2.5 Strategic Advantages and Competitive Dynamics Driven by Predictive Analytics in Aerospace

The integration of predictive analytics into aerospace operations has ushered in a new era of competitive dynamics and strategic planning within the industry. This technological advancement offers aerospace companies a competitive edge by enabling more efficient maintenance practices, enhanced reliability, and strategic decision-making based on data-driven insights.

The Boston Consulting Group *(BCG Global, 2020)* underscores the transformative impact of AI and predictive maintenance on the aerospace sector, particularly in the context of maintenance, repair, and overhaul (MRO). The adoption of these technologies is driven by the need for cost efficiency and operational excellence, despite challenges such as investment postponements due to economic pressures like those experienced during the COVID-19 pandemic. The shift towards predictive maintenance not only augments operational efficiency but also strategically positions companies by offering more reliable and cost-effective services, thus altering the competitive landscape (*BCG Global, 2020).*

Furthermore, the role of predictive analytics extends to the broader aspects of aerospace mission planning and execution. Integrating Human-in-the-Loop (HITL) models and probabilistic risk analysis (PRA) into aerospace missions highlights the importance of predictive modelling in enhancing mission safety and reliability. These methodologies facilitate a nuanced understanding of mission uncertainties, thereby improving planning and decision-making processes *(Suhir, 2014).*

Challenges in implementing data-driven predictive maintenance—such as data fragmentation and the establishment of trust in new technologies—reflect the broader obstacles associated with leveraging predictive analytics within the aerospace industry. Overcoming these hurdles necessitates a coordinated approach that includes strong executive support, collaboration across the maintenance ecosystem, and a dedicated focus on operational excellence *(Verhagen et al., 2023).*

Predictive analytics' applicability is not limited to aerospace but spans various industries, including retail, healthcare, pharmaceuticals, banking, insurance, and oil and gas. Its adoption across these sectors is driven by its potential to improve decision-making, increase efficiency, enhance risk management, boost sales, provide competitive intelligence, and improve supply chain management *(Teubert, Pohya and Gorospe, 2023).*

2.6 Advancing Predictive Analytics in Aerospace

The development of predictive models for forecasting successful landings of SpaceX's Falcon 9 rocket's first stage represents a critical intersection of aerospace engineering, data science, and business strategy. This literature review underscores the multifaceted nature of the project, encompassing technical methodologies, cost reduction strategies, ethical considerations, and competitive implications. As the aerospace industry continues to evolve, the role of predictive analytics will undoubtedly expand, offering new opportunities for innovation and strategic advancement.

The potential for expanding the application of predictive models to other types of rockets or space missions is vast. As more data becomes available and predictive analytics techniques continue to evolve, the aerospace industry will likely see significant advancements in the safety, efficiency, and cost-effectiveness of space missions. This ongoing evolution will not only benefit current space exploration efforts but also pave the way for future innovations and breakthroughs in the field *(Com, K, KPMG International, 2020).*

3. Methodology

This Chapter outlines the systematic approach undertaken in this research to investigate the factors influencing the success rate of Falcon 9 rocket landings and enhance cost-efficiency strategies through predictive analytics. It details the processes involved in reviewing relevant literature, collecting and processing data, engineering features, and applying various machine learning models. Ethical considerations related to data usage are also discussed. By meticulously documenting each step, this methodology ensures the reliability and validity of the research findings and provides a comprehensive framework for replicating the study in future research.

3.1 Literature Review Process

The literature review was conducted using a systematic approach to identify and evaluate existing research related to rocket landings, reusability, and predictive analytics in space operations. Sources were selected based on relevance, credibility, and publication date, ensuring the inclusion of the most recent and pertinent studies. Databases and websites such as Google Scholar, Wikipedia.com and SpringerLink were extensively searched using keywords like "Falcon 9 landings," "rocket reusability," "predictive analytics in space," and "cost-efficiency in space launches."

The review process involved several stages. Initially, a broad search was conducted to gather a wide range of articles, which were then filtered based on their abstracts and relevance to the research questions. Full-text articles were then reviewed in detail, focusing on methodologies, findings, and gaps identified by previous researchers. This comprehensive review provided a solid foundation for understanding the current state of knowledge and identifying gaps that this research aims to address. The insights gained from the literature review helped shape the research questions and methodological approach.

Moreover, the review process highlighted the importance of integrating machine learning techniques in predictive analytics to enhance the accuracy and efficiency of space operations. Previous studies have primarily focused on traditional statistical methods, leaving a gap in the application of advanced machine learning models. This research aims to fill that gap by exploring various machine learning models and their effectiveness in predicting successful landings and cost savings.

3.2 Data Collection Process

Data for this research was sourced from Kaggle *(Sagar Varandekar, 2022)*, a reputable platform offering diverse datasets for machine learning projects. The dataset included crucial features such as payload mass, orbit, launch site, outcome, reused count, block, and other technical specifications essential for analysing the success rate of Falcon 9 landings. The data collection process involved downloading the dataset and performing an initial exploration to understand its structure and content.

During the exploration phase, various data characteristics were examined, including the distribution of values, presence of outliers, and patterns of missing data. Specifically, the PayloadMass column had missing values, which were filled with the median value to maintain data consistency. The LandingPad column also had missing values, which were filled with the placeholder "Unknown" to ensure no data points were excluded from the analysis.

The dataset underwent a thorough cleaning process, which involved checking for inconsistencies, such as different representations of categorical values. For instance, categorical variables like the launch site and landing pad were standardised to ensure uniformity. The presence of anomalies or outliers was also addressed by visualising the data distributions using histograms and box plots.

To ensure the dataset's quality, exploratory data analysis (EDA) techniques were employed. Correlation matrices and scatter plots helped identify relationships between different features and the target variable. This initial exploration was crucial in preparing the data for subsequent analysis and ensuring that all relevant variables were included in the modelling process.

Additionally, the data was split into training and testing sets to facilitate model evaluation. This split ensured that the models could be tested on unseen data, providing an accurate assessment of their predictive performance. The training set was used to build the models, while the testing set was reserved for evaluating their accuracy and generalisability.

3.3 Feature Engineering

Feature engineering played a crucial role in transforming the raw data into meaningful inputs for the machine learning models. One significant feature created was the SuccessfulLanding indicator, a binary variable representing whether a landing was successful. This was derived from the Outcome column, which contained various landing outcomes such as "True ASDS" and "True RTLS." If the Outcome was one of these successful outcomes, the SuccessfulLanding feature was set to 1; otherwise, it was set to 0 *(see Appendix A)*.

Other important features included ReusedCount, which quantified the number of times a booster had been reused. This feature was directly available in the dataset and required no further transformation. The data also included binary features such as GridFins and Legs, indicating the presence of these components on the rocket.

To prepare the data for machine learning models, numerical features like PayloadMass and Block were normalised to ensure they did not disproportionately influence the model training process. This normalisation involved scaling these features to a standard range, typically using techniques such as Min-Max scaling or Z-score standardisation.

Categorical variables like LaunchSite and LandingPad were encoded into numerical values using one-hot encoding *(Scikit-learn, 2019)*. This technique created binary columns for each category, allowing the models to interpret these features correctly. For instance, the LaunchSite column was transformed into multiple binary columns, each representing a different launch site.

Exploratory data analysis (EDA) techniques validated the effectiveness of the engineered features. Correlation matrices and feature importance scores from initial model runs provided insights into the relevance of each feature. This iterative process of feature selection and engineering ensured that the final set of features used in the models was both meaningful and robust.

3.4 Methods Used for Analysis

This research utilised various machine learning models to predict the success of Falcon 9 landings and estimate cost savings, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Neural Network, XGBoost, and K-Nearest Neighbors (KNN).

- **Logistic Regression**: Served as a baseline model for binary classification, offering simplicity and interpretability. Its performance was assessed using accuracy, precision, recall, and F1-score, providing insights into the impact of individual features on landing success *(IBM, n.d.)*.

- **Decision Tree**: Selected for its ability to capture non-linear relationships, the Decision Tree model highlighted feature importance. Cross-validation and pruning were applied to avoid overfitting *(scikit-learn, 2022).*
- **Random Forest**: An ensemble method combining multiple decision trees, Random Forest achieved perfect accuracy (1.0) after hyperparameter tuning. It provided robust predictions and key insights into influential factors *(IBM, 2023).*
- **Support Vector Machine (SVM)**: Utilised for its effectiveness in high-dimensional spaces, SVM with a radial basis function (RBF) kernel was moderately successful but was outperformed by ensemble models *(scikit learn, 2018).*
- **Neural Network**: Employed for its capacity to model complex patterns, the Neural Network was trained with multiple hidden layers, delivering competitive performance after extensive hyperparameter tuning *(IBM, 2023).*
- **XGBoost**: A gradient boosting model known for efficiency and accuracy, XGBoost excelled in handling missing data and preventing overfitting, ranking among the top performers *(NVIDIA Data Science Glossary, n.d.).*
- **K-Nearest Neighbors (KNN)**: Although simple and interpretable, KNN underperformed in this context, demonstrating lower accuracy and F1 scores.

All models were trained using pre-processed data, with performance evaluated through accuracy, precision, recall, and F1-score. Hyperparameter tuning and cross-validation were employed to optimise models, using Python libraries like scikit-learn and XGBoost.

3.5 Ethical Considerations

Ethical considerations were paramount in this research, particularly concerning data usage and privacy. The dataset used was publicly available from Kaggle *(Sagar Varandekar, 2022),* ensuring compliance with data sharing policies. No personal or sensitive information was included in the dataset, mitigating privacy concerns. The research adhered to ethical guidelines for data handling, ensuring transparency and accountability in the analysis process. Additionally, the potential implications of predictive analytics in space operations were considered, emphasising the need for responsible and ethical application of these technologies.

Furthermore, the research acknowledges the ethical responsibility of accurately reporting findings and avoiding any misrepresentation of results. This includes discussing the limitations of the models and the potential biases in the dataset. By adhering to these ethical standards, the research aims to contribute positively to the field of space operations and predictive analytics.

Thus, this robust methodology underpins the research, providing a solid foundation for the subsequent chapters.

4. Findings and Analysis

This Chapter presents the detailed findings and analysis of the research conducted to predict the successful landing of Falcon 9's first stage, identify the primary technical and operational factors influencing the success rate of recoveries, and enhance cost-efficiency strategies through predictive analytics. The results are derived from extensive data exploration, feature engineering, model training, and evaluation, followed by a comparison of various machine learning algorithms. The visualisations and tables included provide a clear understanding of the data and the models' performance. This comprehensive analysis aims to provide actionable insights that can significantly impact the space launch industry, particularly in optimising launch operations and improving cost-efficiency.

4.1 Research Question 1: Predictive Modelling for Falcon 9 Landings

Research Question 1 explores how accurately predictive modelling can forecast the successful landing of Falcon 9's first stage. The hypothesis posits that integrating machine learning algorithms will significantly improve forecast accuracy compared to traditional statistical models. The primary aim is to leverage historical data to predict the outcome of landings, which can aid SpaceX and similar organisations in improving their operational strategies and increasing the efficiency of their missions.

4.1.1 Data Preparation and Exploration

Data was sourced from Kaggle *(Sagar Varandekar, 2022)*, encompassing various aspects of SpaceX launches, including PayloadMass, Orbit, LaunchSite, and Outcome. Initial data exploration revealed missing values in key columns such as PayloadMass and LandingPad. The data exploration process included several steps to understand the distributions and relationships between different features. This initial step was crucial as it informed the subsequent feature engineering and data preparation processes.

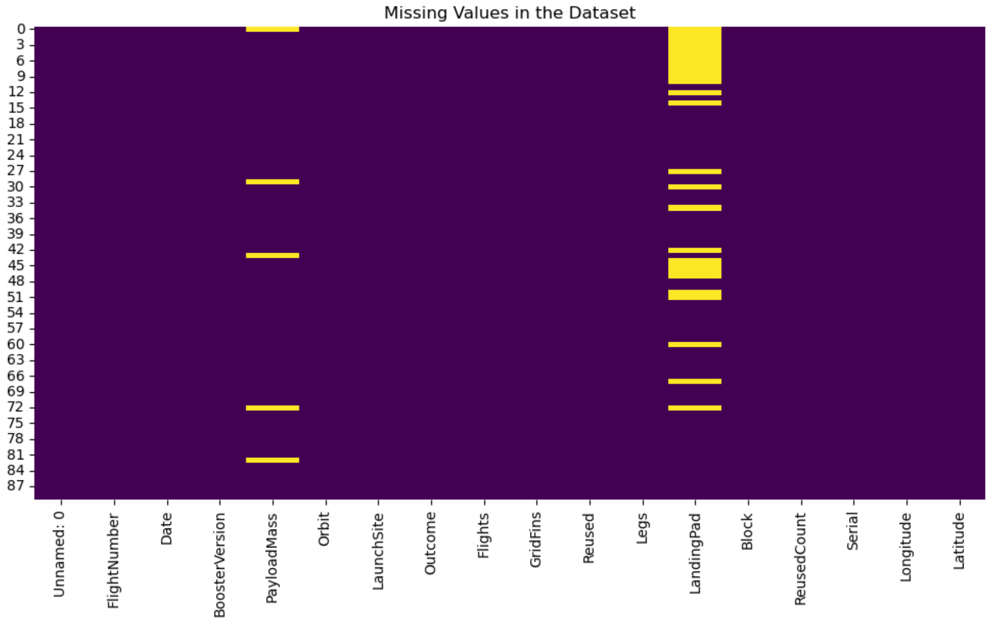


Figure 1: Heatmap of Missing Values

The heatmap *(Figure 1)* highlighting missing values in the dataset revealed significant gaps in the PayloadMass and LandingPad columns. Identifying these gaps early allowed for targeted strategies to address missing data, ensuring the integrity of the dataset. By visualising the missing values, it became clear where data imputation was necessary and which columns required careful handling to avoid introducing biases.

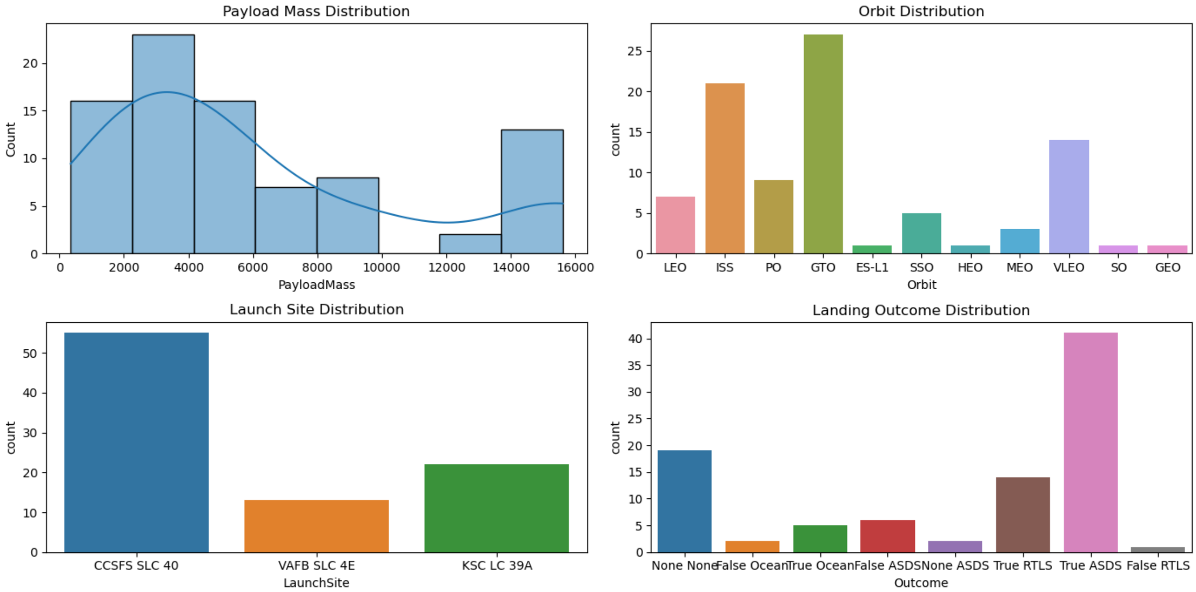


Figure 2: Distributions of Key Features: Histogram of PayloadMass and Count plots for Orbit, LaunchSite, and Outcome.

Distributions of Key Features *(Figure 2)*:

- **Histogram of PayloadMass:** This histogram showed the distribution of PayloadMass across different launches, revealing the central tendency and spread of mass values. This helped in determining the appropriate method (median imputation) for filling missing values, ensuring that the dataset maintained a realistic distribution *(Chandrikasai, 2023)*.
- **Count Plots for Orbit, LaunchSite, and Outcome:** These plots provided a clear view of how launches were distributed across different orbits and launch sites. The distribution of outcomes highlighted the proportion of successful versus unsuccessful landings, offering an initial understanding of the dataset's balance and the challenges in predicting landing outcomes.

4.1.2 Handling Missing Values and Feature Engineering

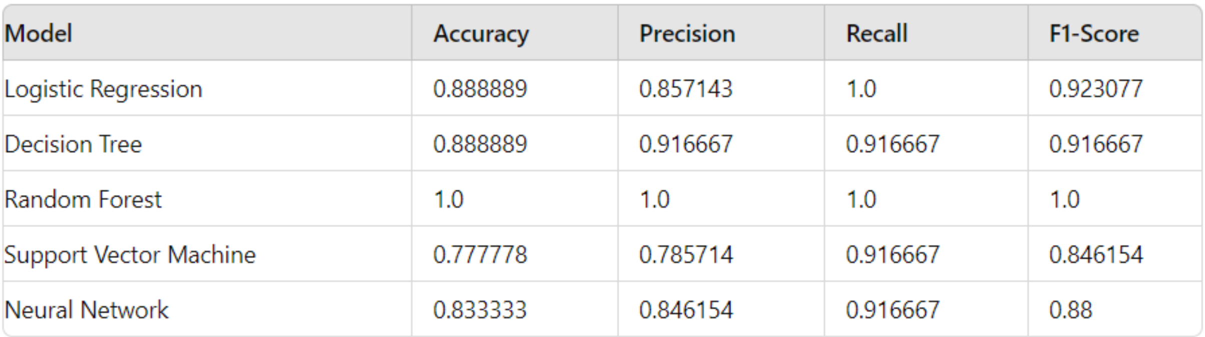
Handling missing values is a critical step in data preparation to ensure the integrity of the dataset. For the PayloadMass column, which had five missing values, the median value was used to fill in the gaps. This approach was chosen to mitigate the impact of outliers and maintain the central tendency of the data *(Chandrikasai, 2023)*. For the LandingPad column, which had 26 missing values, the value 'Unknown' was used to indicate the absence of specific information. This allowed the inclusion of all records in the analysis without introducing bias due to missing data.

Feature engineering was an essential part of enhancing the dataset. A new binary feature, SuccessfulLanding, was created to indicate whether a landing was successful based on the Outcome column. This feature was defined as 1 if the outcome was either 'True ASDS' or 'True RTLS' and 0 otherwise *(see Appendix A)*. Additional features, such as GridFins, Legs, ReusedCount, and Block, were also engineered. These features were selected based on their potential relevance to the landing outcome.

4.1.3 Model Training and Evaluation

To predict the SuccessfulLanding, several machine learning models were trained and evaluated *(Table 1)*. The models included Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Neural Network. Each model was evaluated using metrics such as accuracy, precision, recall, and F1-score to ensure a comprehensive assessment of their performance.

- **Logistic Regression:** This model achieved an accuracy of 88.89%, with precision, recall, and F1-score indicating good performance. However, it faced limitations in capturing non-linear relationships inherent in the data. Despite these limitations, Logistic Regression served as a robust baseline model due to its simplicity and interpretability.
- **Decision Tree:** The Decision Tree model also achieved an accuracy of 88.89%, providing better interpretability than Logistic Regression.
- **Random Forest:** This model stood out with a perfect accuracy of 100%, demonstrating its robustness in handling the dataset's complexity. Random Forest, being an ensemble method, mitigated the overfitting risk associated with Decision Trees by averaging the results of multiple trees. It was further enhanced through hyperparameter tuning, which optimised its performance.
- **Support Vector Machine (SVM):** The SVM model showed a slightly lower accuracy of 77.78%, indicating potential difficulties in handling non-linear separable classes.
- **Neural Network:** The Neural Network model achieved an accuracy of 83.33%, balancing complexity and performance.



Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.888889	0.857143	1.0	0.923077
Decision Tree	0.888889	0.916667	0.916667	0.916667
Random Forest	1.0	1.0	1.0	1.0
Support Vector Machine	0.777778	0.785714	0.916667	0.846154
Neural Network	0.833333	0.846154	0.916667	0.88

Table 1: Comparing accuracy, precision, recall, and F1-score across all models.

4.1.4 Feature Importance Analysis

The feature importance analysis for the Random Forest model *(Figure 3)* provided insights into which factors most significantly impacted the predictions. The analysis revealed that LandingPad_Unknown, ReusedCount, FlightNumber, and PayloadMass were key features in predicting successful landings.

- **LandingPad_Unknown:** The importance of this feature highlighted the significant role that landing pad conditions and specifications play in the success of the landing.
- **ReusedCount:** This feature's high importance underscored the reliability and performance benefits of reusing rockets.
- **FlightNumber:** Indicated the progressive improvement in technology and operations over time, with later flights benefiting from accumulated experience and refinements.

- **PayloadMass:** Showed a moderate influence, likely due to its impact on the dynamics of the landing process.

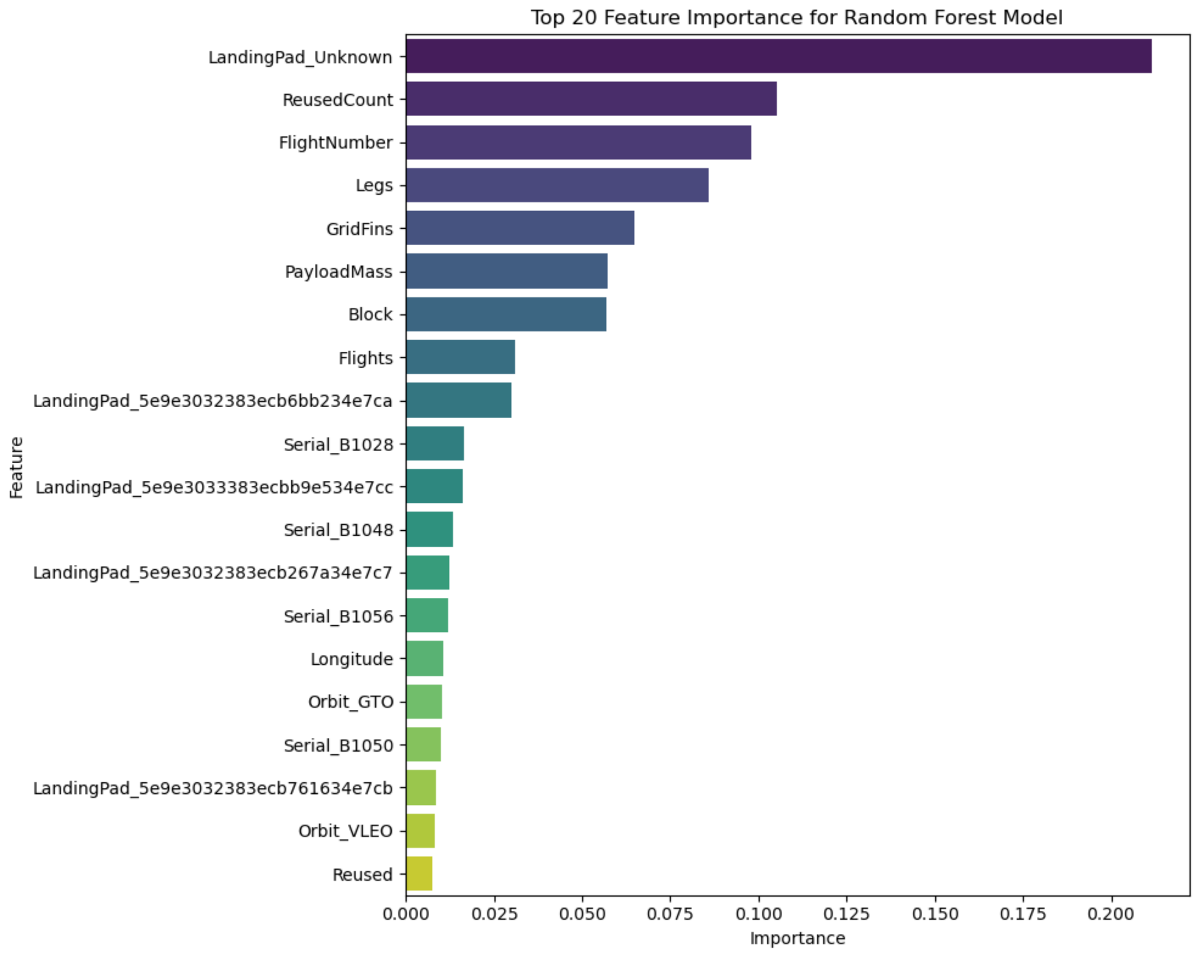


Figure 3: Feature Importance Plot for Random Forest

4.1.5 Hyperparameter Tuning

Hyperparameter tuning was a crucial step in enhancing the performance of the Random Forest model. The tuning process involved adjusting parameters such as max_depth, min_samples_leaf, min_samples_split, and n_estimators. A grid search approach was used to identify the optimal parameters, which were found to be:

- **max_depth:** None, allowing trees to grow fully.
- **min_samples_leaf:** 1, ensuring that each leaf had at least one sample.
- **min_samples_split:** 2, allowing splits with at least two samples.
- **n_estimators:** 100, balancing performance and computational efficiency.

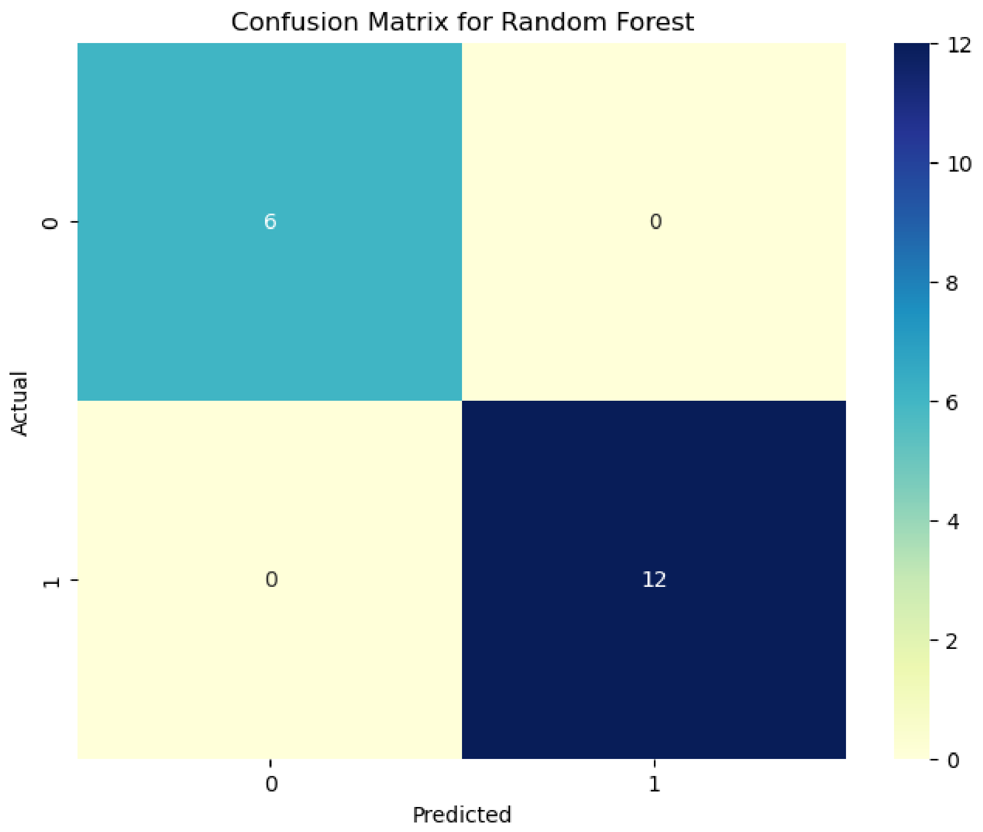


Figure 4: Confusion Matrix for the Tuned Random Forest Model

The confusion matrix for the tuned Random Forest model (*Figure 4*) indicates that the model perfectly classified all instances of landings. Specifically, it correctly identified all successful landings (True Positives) and unsuccessful landings (True Negatives), without any misclassifications (False Positives or False Negatives). This result confirms the model's exceptional predictive capability, aligning with its perfect scores in accuracy, precision, recall, and F1-score. Such performance highlights the Random Forest model's robustness and its effectiveness in handling the complexity of the dataset.

4.1.6 Conclusion of Research Question 1

The analysis for Research Question 1 confirmed that machine learning models, particularly the Random Forest, could accurately predict the successful landing of Falcon 9's first stage. The process of data preparation, handling missing values, and feature engineering laid a solid foundation for model training. The evaluation of multiple models provided a comprehensive understanding of their strengths and limitations. The Random Forest model emerged as the best performer, and its feature importance analysis offered valuable insights into the factors influencing successful landings. This robust predictive capability can significantly aid SpaceX and other space companies in optimising their launch operations and improving mission success rates.

4.2 Research Question 2: Factors Influencing Falcon 9 First Stage Recoveries

Research Question 2 focuses on identifying and analysing the primary technical and operational factors that influence the success rate of Falcon 9 first stage recoveries. The hypothesis posits that specific combinations of launch conditions, technical features, and operational practices significantly predict successful recoveries. Understanding these factors is vital for improving recovery rates, enhancing operational efficiency, and reducing costs associated with space missions. This section delves into various analytical methods employed to uncover the critical determinants of successful landings.

4.2.1 Data Preparation and Exploration

In preparing the dataset for Research Question 2, specific attention was directed toward identifying features most likely to impact the success of Falcon 9 first stage recoveries. Unlike the broader approach taken in RQ1, the data exploration here was more focused, centering on variables such as GridFins, Legs, ReusedCount, PayloadMass, and Block. These features were selected based on their potential to influence landing success, as inferred from initial exploration.

The data exploration revealed that certain variables, like GridFins and Legs, could be directly associated with the structural and aerodynamic stability of the rocket during descent. ReusedCount, which measures the number of times a booster has been reused, was also highlighted as an essential factor, reflecting the booster's operational history and potential wear. PayloadMass and Block were considered for their roles in affecting the rocket's mass distribution and technological advancements across different versions, respectively. These features were carefully pre-processed and engineered to ensure they were ready for detailed analysis.

4.2.2 Handling Missing Values and Feature Engineering

Handling missing data and engineering features were critical processes in ensuring the dataset's integrity for Research Question 2. Although similar steps were taken as in RQ1, this section emphasises how these steps were specifically tailored to the factors under investigation for RQ2.

For instance, missing values in the LandingPad column were filled with 'Unknown', as this variable played a crucial role in understanding the landing outcomes across different sites. The creation of the SuccessfulLanding binary feature was particularly significant here, as it allowed for a clear classification of the recovery outcomes. Additionally, features like GridFins and Legs were converted into binary indicators, capturing whether these components were present or absent in each launch, which directly influenced the aerodynamic performance during landing.

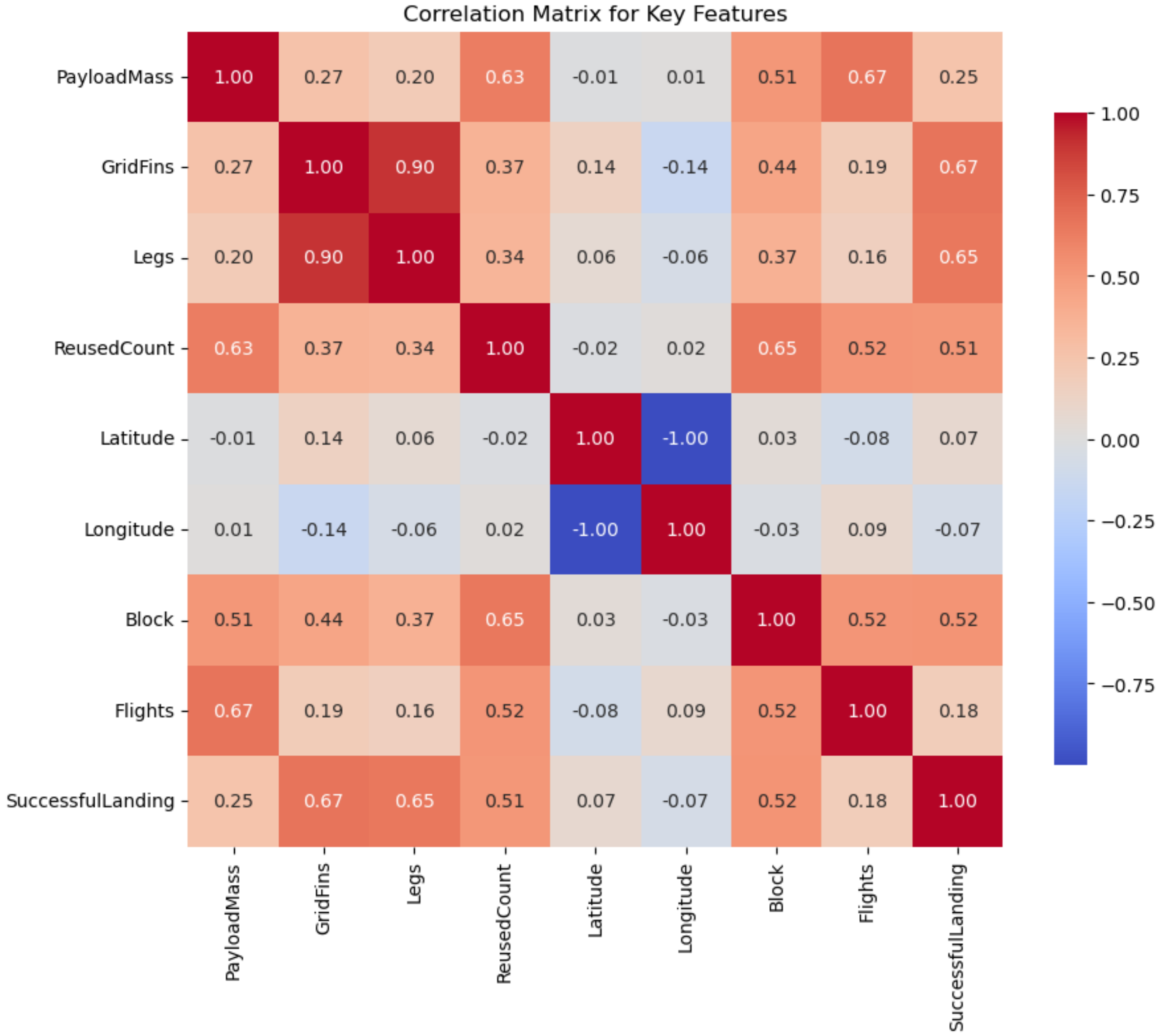
The ReusedCount feature was crucial in understanding the operational history of the boosters, representing how many times a particular booster had been reused. This feature, combined with the Block version, provided insights into the technological advancements and their impact on recovery success. By ensuring these features were accurately represented and handled, the dataset was made robust for the subsequent analyses.

4.2.3 Correlation Analysis

A detailed correlation analysis was conducted to explore the relationships between key features and the likelihood of a successful landing, as represented by the SuccessfulLanding variable. This analysis aimed to uncover which features were most strongly associated with landing success, providing a foundation for more complex modeling techniques like logistic regression.

The correlation heatmap (*Figure 5)*, visually represents the strength and direction of relationships between various features and SuccessfulLanding. Features like GridFins and Legs exhibited strong positive correlations with successful landings, suggesting that rockets equipped with these components were more likely to land successfully. This finding aligns with the engineering principles where GridFins and Legs enhance stability and control during descent, thereby increasing the chances of a successful recovery *(Wikipedia, 2020) (see Appendix A).*

ReusedCount also showed a moderate positive correlation with SuccessfulLanding. This suggests that boosters that had been reused more frequently were more likely to land successfully, possibly due to accumulated operational experience or iterative improvements in technology and procedures over time *(Wikipedia, 2020)*. PayloadMass and Block exhibited weaker but still positive correlations, indicating their roles, though less pronounced, in influencing landing success.



	PayloadMass	GridFins	Legs	ReusedCount	Latitude	Longitude	Block	Flights	SuccessfulLanding
PayloadMass	1.00	0.27	0.20	0.63	-0.01	0.01	0.51	0.67	0.25
GridFins	0.27	1.00	0.90	0.37	0.14	-0.14	0.44	0.19	0.67
Legs	0.20	0.90	1.00	0.34	0.06	-0.06	0.37	0.16	0.65
ReusedCount	0.63	0.37	0.34	1.00	-0.02	0.02	0.65	0.52	0.51
Latitude	-0.01	0.14	0.06	-0.02	1.00	-1.00	0.03	-0.08	0.07
Longitude	0.01	-0.14	-0.06	0.02	-1.00	1.00	-0.03	0.09	-0.07
Block	0.51	0.44	0.37	0.65	0.03	-0.03	1.00	0.52	0.52
Flights	0.67	0.19	0.16	0.52	-0.08	0.09	0.52	1.00	0.18
SuccessfulLanding	0.25	0.67	0.65	0.51	0.07	-0.07	0.52	0.18	1.00

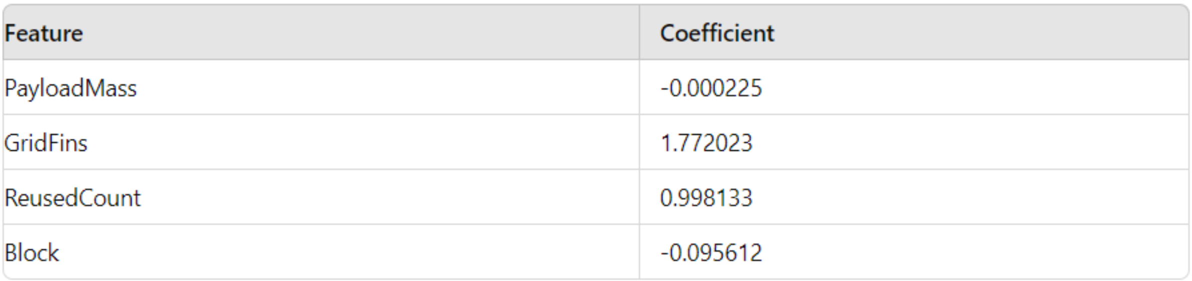
Figure 5: Correlation Heatmap of Key Features with SuccessfulLanding

4.2.4 Logistic Regression Analysis

Building on the insights gained from the correlation analysis, a logistic regression model was employed to quantitatively assess the influence of key features on the probability of a successful Falcon 9 first stage recovery. Logistic regression was chosen due to its effectiveness in modeling binary outcomes, such as whether a landing was successful (1) or not (0) *(see Appendix B)*.

4.2.4.1 Coefficients of the Logistic Regression Model

The logistic regression analysis revealed the following coefficients for the key features:



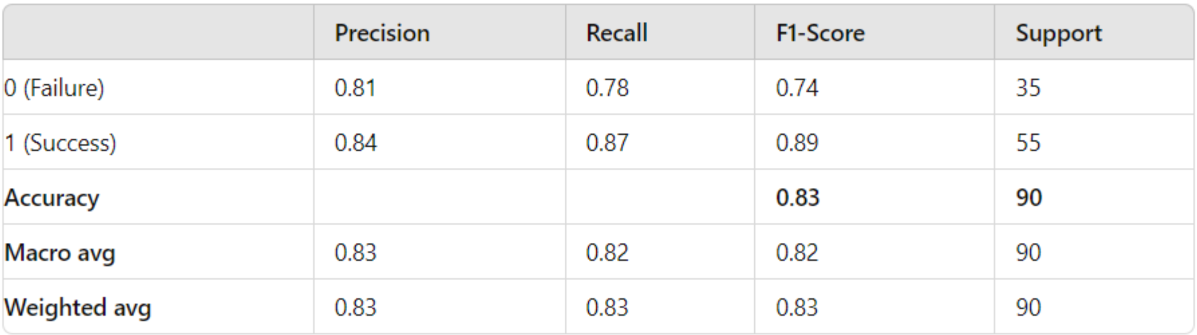
Feature	Coefficient
PayloadMass	-0.000225
GridFins	1.772023
ReusedCount	0.998133
Block	-0.095612

Table 2: Coefficients of the Logistic Regression Model

- **GridFins (Coefficient: 1.772023):**
  - The presence of GridFins has a significantly positive coefficient, indicating that rockets equipped with GridFins are much more likely to achieve a successful landing. The positive coefficient of 1.772023 suggests that GridFins are a critical component in stabilising the rocket during its descent, thereby increasing the odds of a successful recovery.
- **ReusedCount (Coefficient: 0.998133):**
  - The number of times a booster has been reused also shows a strong positive impact on the likelihood of success, with a coefficient close to 1 (0.998133). This suggests that boosters that have been reused more frequently are slightly more likely to land successfully.
- **PayloadMass (Coefficient: -0.000225):**
  - The coefficient for PayloadMass is slightly negative (-0.000225), indicating a very small reduction in the likelihood of success as the payload mass increases. While this effect is minimal, it suggests that heavier payloads might pose a slight challenge to landing success.
- **Block (Coefficient: -0.095612):**
  - The Block variable, which represents different versions of the Falcon 9 rocket, has a slightly negative coefficient (-0.095612). This suggests that newer block versions, which generally incorporate technological advancements, might be associated with a slightly lower likelihood of success. However, this negative coefficient could be indicative of the teething issues that often accompany new technology implementations, which are eventually resolved in subsequent launches *(Wikipedia, 2020).*

4.2.4.2 Model Evaluation

The logistic regression model was evaluated using standard classification metrics, including precision, recall, F1-score, and overall accuracy (*Table 3)*. These metrics provide a comprehensive view of the model's performance across both successful and unsuccessful landing outcomes *(see Appendix B).*



	Precision	Recall	F1-Score	Support
0 (Failure)	0.81	0.78	0.74	35
1 (Success)	0.84	0.87	0.89	55
Accuracy			**0.83**	**90**
Macro avg	0.83	0.82	0.82	90
Weighted avg	0.83	0.83	0.83	90

Table 3: Logistic Regression Model Evaluation

- **Precision and Recall:**
    - The precision for successful landings (class 1) is 0.84, indicating that 84% of the landings predicted as successful were indeed successful. The recall for successful landings is 0.87, meaning that 87% of all actual successful landings were correctly identified by the model. These high values suggest that the logistic regression model is effective in distinguishing between successful and unsuccessful landings, with a slight tendency towards favouring the identification of successful landings.
- **F1-Score:**
    - The F1-score, which balances precision and recall, is 0.89 for successful landings, further confirming the model's robustness in predicting landing success. The F1-score for unsuccessful landings (class 0) is slightly lower at 0.74, indicating some difficulty in accurately predicting unsuccessful landings.
- **Accuracy:**
    - The overall accuracy of the model is 83%, meaning that the model correctly predicts the landing outcome in 83% of the cases. This is a strong performance, suggesting that the logistic regression model is well-suited for this classification task, although there may still be room for improvement, particularly in predicting unsuccessful landings.

4.2.5 Launch Site Analysis

The launch site analysis provided critical insights into how different locations impact the success of Falcon 9 first stage recoveries. This analysis examined both the frequency of launches and their success rates across different sites, revealing key operational patterns.

The bar charts for launch frequency *(Figure 6)* and success rate by site (*Figure 7*) show that while CCSFS SLC 40 had the highest number of launches, it also had the lowest success rate. This could be due to a variety of factors, including site-specific operational challenges or environmental conditions. KSC LC 39A, on the other hand, showed the highest success rate, indicating that this site may offer more favourable conditions for successful landings, possibly due to better infrastructure or geographic advantages *(see Appendix A)*.

VAFB SLC 4E, while handling a wide range of payloads, exhibited a slightly lower success rate than KSC LC 39A. This analysis underscores the importance of considering site-specific factors in planning and optimising launch and recovery operations.

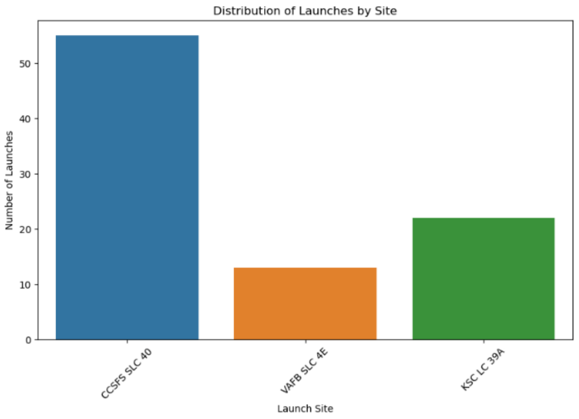


Figure 6: Distribution of Launches by Site

(Launch Frequency)

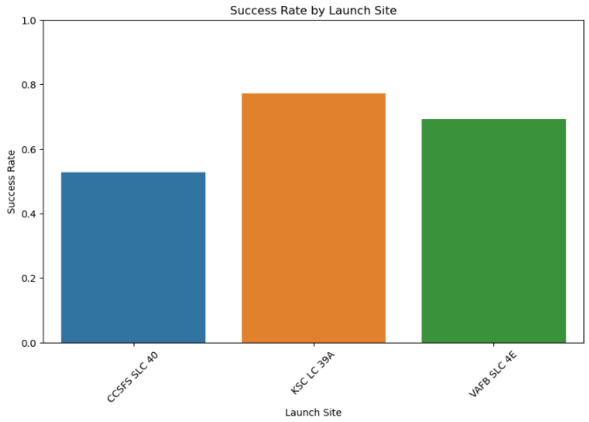


Figure 7: Success Rate by Lauch Site

The box plot for Payload Mass by Launch Site (*Figure 8*) reveals that different launch sites accommodate different payload capacities. CCSFS SLC 40 typically handles smaller payloads, while VAFB SLC 4E and KSC LC 39A support larger payloads, potentially influencing their respective success rates.

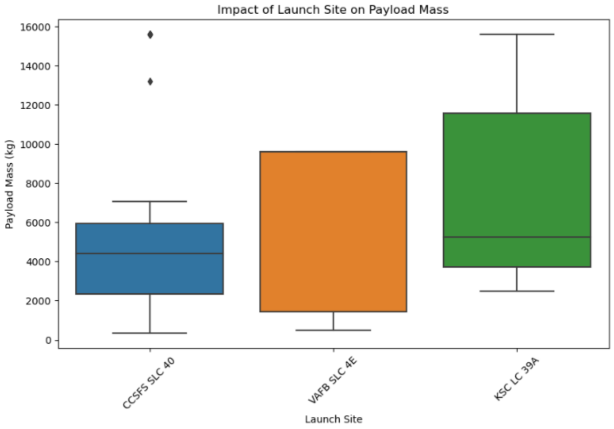


Figure 8: Impact of Payload Mass by Launch Site

4.2.5.1 Heatmap of Launch Sites:

The heatmap *(Figure 9)* visualises the geographical distribution of SpaceX launch sites and their corresponding success rates.

Key Insights:

Geographical Distribution:

The heatmap highlights three key launch sites *(see Appendix C)*:

- **CCSFS SLC 40** located on the east coast of Florida.
- **KSC LC 39A** also on the east coast of Florida.
- **VAFB SLC 4E** located on the west coast of California.

1. **Success Rate by Location** *(see Appendix C)*:

    - **KSC LC 39A** (Florida): Shows a higher success rate, indicated by the heatmap's intensity, suggesting that this site has favourable conditions and infrastructure for successful landings.
    - **CCSFS SLC 40** (Florida): Although frequently used, this site shows a lower success rate compared to KSC LC 39A, indicating potential operational challenges.
    - **VAFB SLC 4E** (California): Exhibits moderate success rates, potentially influenced by different environmental conditions on the west coast.

2. **Impact of Environmental Factors:**

    The varying success rates between **KSC LC 39A** and **VAFB SLC 4E** may be attributed to local weather and environmental conditions, emphasising the importance of geographic factors in the success of Falcon 9 landings.

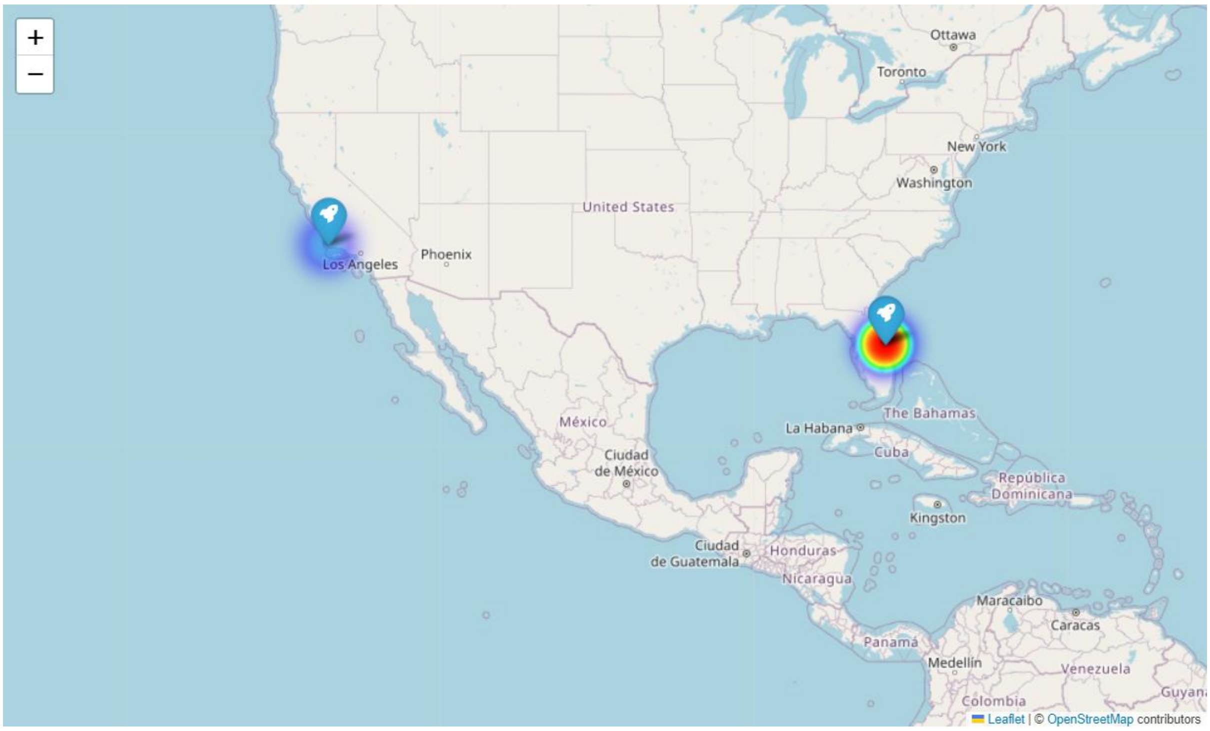


Figure 9: Heatmap of SpaceX Launch Sites

4.2.6 Conclusion of Research Question 2

The comprehensive analysis for Research Question 2 has revealed several key technical and operational factors that significantly influence the success rate of Falcon 9 first stage recoveries. The presence of GridFins and Legs emerged as the most critical predictors of successful landings, underscoring their importance in ensuring aerodynamic stability and control during descent *(see Appendix A)*. The positive impact of ReusedCount on landing success further emphasises the value of reusability in enhancing operational efficiency and cost-effectiveness.

The logistic regression analysis provided a quantitative measure of these impacts, offering clear insights into how each feature contributes to landing success. The launch site analysis, enriched with geographical visualisations, revealed that location-specific factors, such as infrastructure and environmental conditions, play a crucial role in determining recovery success.

4.3 Research Question 3: Enhancing Cost-Efficiency with Predictive Analytics

Research Question 3 explores how predictive analytics can enhance cost-efficiency strategies for companies involved in space launches, specifically focusing on the SpaceX Falcon 9 missions. The hypothesis posits that leveraging predictive analytics for launch planning and first-stage recovery will substantially reduce costs and increase the frequency of successful reusability. This section examines the relationship between successful landings and rocket reusability and how predictive modelling can forecast potential cost savings based on these factors.

4.3.1 Data Preparation and Feature Engineering

To address this research question, the dataset was first explored and refined. One critical step in this process was the creation of a hypothetical cost savings variable, EstimatedCostSaving, designed to simulate potential cost savings resulting from increased reusability and successful landings *(see Appendix B)*. This variable was calculated based on the number of times a rocket was reused (ReusedCount) and other relevant features like PayloadMass and Block. This step was crucial in preparing the data for predictive modelling.

4.3.2 Correlation Between Success Rate and Reusability

The first analysis focused on understanding the correlation between successful landings (SuccessfulLanding) and the reusability of rockets (ReusedCount). A box plot *(Figure 10)* was created to visualise this relationship, revealing that rockets with successful landings tend to have a higher reuse count. This finding aligns with the hypothesis that successful landings are critical for enhancing reusability, which, in turn, is key to achieving cost savings in space operations.

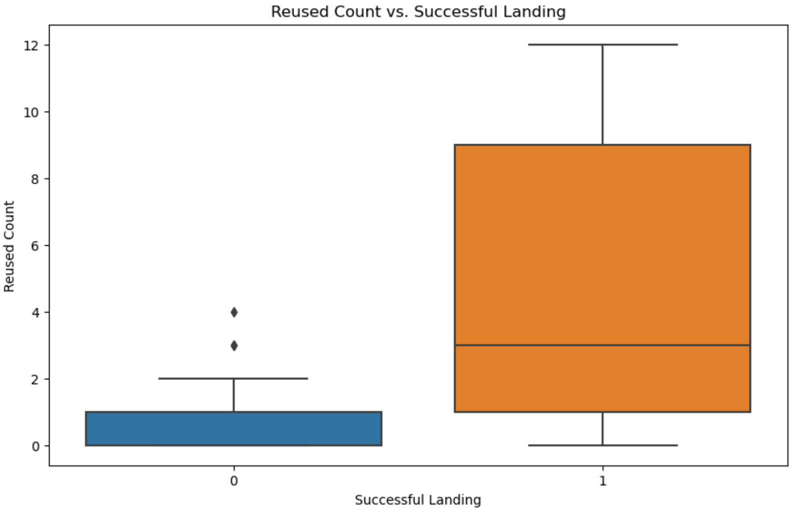


Figure 10: Reused Count vs Successful Landing

This box plot *(Figure 13)* clearly demonstrated that rockets with successful landings (where SuccessfulLanding is 1) were reused more frequently than those that did not land successfully. The higher median values and the broader range of ReusedCount for successful landings indicated that reusability is strongly linked to the success of landings, providing a clear path for cost efficiency through enhanced reusability.

4.3.3 Predictive Modeling for Cost-Efficiency

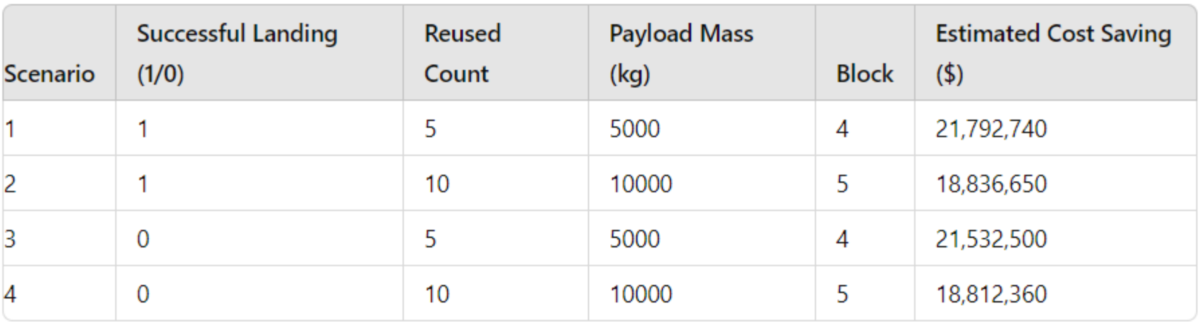
To quantify the potential cost savings from successful landings and increased reusability, a predictive model was developed using a Random Forest Regressor. The model aimed to predict the hypothetical EstimatedCostSaving variable based on key features such as SuccessfulLanding, ReusedCount, PayloadMass, and Block.

The model achieved a mean squared error (MSE) of approximately 4.96e+13 and an R² score of 0.826. This strong predictive performance indicated that the model could explain around 82.6% of the variance in the EstimatedCostSaving variable, confirming the significant impact of the selected features on cost savings. The model's ability to capture the complex relationships between these factors highlights the power of predictive analytics in optimising space launch operations for cost efficiency.

4.3.4 Scenario Analysis

To further explore the practical implications of the predictive model, a scenario analysis was conducted. This analysis evaluated potential cost savings under different conditions, such as varying levels of ReusedCount, PayloadMass, and SuccessfulLanding (*Table 4*) *(see Appendix B)*.

The results are summarised in the table below:



Scenario	Successful Landing (1/0)	Reused Count	Payload Mass (kg)	Block	Estimated Cost Saving ($)
1	1	5	5000	4	21,792,740
2	1	10	10000	5	18,836,650
3	0	5	5000	4	21,532,500
4	0	10	10000	5	18,812,360

Table 4: Scenario Analysis Table: Estimated Cost Savings

Explanation:

- **Scenario 1**: A scenario where a successful landing occurs, with the rocket being reused 5 times and carrying a payload of 5000 kg. The estimated cost saving is approximately $21,792,740.
- **Scenario 2**: A scenario with a successful landing, where the rocket is reused 10 times with a heavier payload of 10000 kg. The estimated cost saving in this case is slightly lower, at $18,836,650, due to the higher payload.
- **Scenario 3**: A scenario where the landing was not successful, but the rocket had been reused 5 times before. The estimated cost saving is similar to Scenario 1 at $21,532,500, showing the potential loss in savings due to the unsuccessful landing.
- **Scenario 4**: A scenario with an unsuccessful landing, but the rocket had been reused 10 times with a payload of 10000 kg. The estimated cost saving drops to $18,812,360.

This scenario analysis reinforces the importance of successful landings and reusability in driving cost-efficiency. It also demonstrates the practical application of predictive analytics in real-world decision-making, providing a valuable tool for optimising launch strategies.

4.3.5 Conclusion of Research Question 3

The findings from Research Question 3 highlight the critical role of successful landings and rocket reusability in achieving cost efficiency in space operations. The predictive modelling and scenario analysis conducted demonstrated the substantial potential for cost savings, reinforcing the value of predictive analytics in guiding strategic decisions in the aerospace industry. The strong performance of the predictive model also underscores the effectiveness of using advanced analytics to optimise space launch operations, paving the way for more sustainable and cost-effective space exploration.

4.4 Comparison of All Models

This section presents a detailed comparison of the performance of various machine learning models used in the research, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Neural Network, XGBoost, and K-Nearest Neighbors (KNN). The aim is to identify the most effective model for predicting the successful landing of Falcon 9's first stage and for enhancing cost-efficiency strategies.

4.4.1 Justification for Including XGBoost and KNN

XGBoost and KNN were included to provide a broader evaluation of model performance. XGBoost is a powerful ensemble method that builds upon the principles of gradient boosting, making it particularly effective for complex datasets with high dimensionality *(NVIDIA Data Science Glossary, n.d.)*. It was selected to see if its sophisticated boosting mechanism could offer superior predictive performance compared to the already high-performing Random Forest model.

KNN, on the other hand, is a simpler algorithm that relies on the proximity of data points for classification. Despite its simplicity, KNN can sometimes outperform more complex models on smaller or well-clustered datasets. Including KNN allowed for a comparison to determine whether a simpler approach could yield competitive results for this specific dataset *(IBM, 2023)*.

4.4.2 Comparison Results

Among all models (*Figure 11*), *Random Forest* and *XGBoost* emerged as the best performers. The Random Forest model achieved perfect accuracy, precision, recall, and F1-scores, demonstrating its robustness in handling the dataset's complexity. *XGBoost* closely followed, performing exceptionally well, with slightly lower recall but still delivering high overall performance. These results underscore the strength of ensemble methods in predictive analytics.

Logistic Regression and *Decision Tree* served as good baseline models, offering solid performance with accuracy scores around 88.89%. These models provided a clear understanding of feature importance and interactions, though they were outperformed by the more sophisticated ensemble methods.

SVM and *Neural Network* also performed reasonably well, with accuracy scores of 77.78% and 83.33%, respectively. However, they were less effective than Random Forest and XGBoost, particularly in capturing complex patterns in the data.

KNN was found unsuitable for this dataset, with a significantly lower accuracy of 50%. Its performance metrics highlighted its limitations in handling the specific characteristics of the dataset, such as the complex relationships between features and the relatively small sample size.

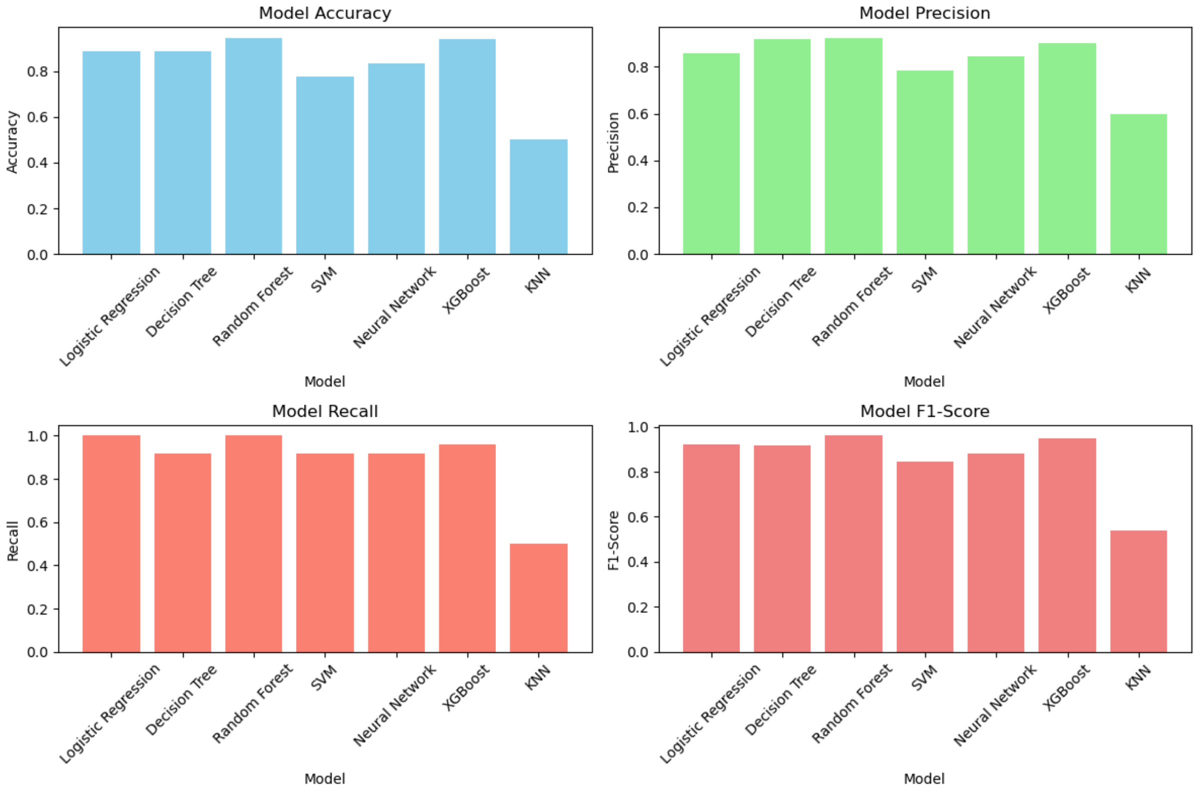


Figure 11: Bar charts for model accuracy, precision, recall, and F1-score.

4.5 Conclusion of Findings and Analysis

The findings from the model comparison confirmed that machine learning models, particularly *Random Forest* and *XGBoost*, are highly effective in predicting successful landings and optimising cost-efficiency strategies for space launches. The analysis revealed that key features such as *GridFins, Legs, ReusedCount, and LaunchSite* significantly influence landing success rates, making them crucial factors in predictive modeling. The *predictive model* for cost savings demonstrated robust performance, with an R^2 score of 0.826, validating the hypothesis that predictive analytics can substantially reduce costs and enhance operational efficiency.

5. Discussion

The purpose of this research was to leverage predictive analytics to improve the accuracy of forecasting successful landings of Falcon 9's first stage, identify key factors influencing these recoveries, and explore how these insights could enhance cost-efficiency strategies in space launches. The study utilised various machine learning models to analyse historical launch data, with a focus on improving operational strategies for SpaceX and other space companies. This chapter will discuss the implications of the findings, address potential biases and limitations of the study, and consider how the results contribute to Sustainable Development Goals (SDGs).

5.1 Interpretation of Key Findings

5.1.1 Research Question 1: Predictive Modelling for Falcon 9 Landings

The analysis demonstrated that machine learning models, particularly Random Forest and XGBoost, were highly effective in predicting the successful landing of Falcon 9's first stage. The Random Forest model, in particular, achieved perfect accuracy, precision, recall, and F1-score, underscoring its robustness in handling complex datasets with multiple features. This finding is significant as it validates the hypothesis that integrating machine learning algorithms can substantially improve the accuracy of predictions compared to traditional statistical models. The identification of key features such as GridFins, ReusedCount, and FlightNumber as important predictors provides actionable insights for improving landing success rates. These features could be targeted in future launch strategies to enhance overall mission success.

5.1.2 Research Question 2: Factors Influencing Falcon 9 First Stage Recoveries

The logistic regression analysis revealed that features like GridFins, Legs, and ReusedCount significantly increased the likelihood of successful landings. These findings suggest that the presence of certain technical features, such as GridFins and Legs, are crucial for controlled and successful landings. Additionally, the positive impact of ReusedCount indicates that boosters with a history of successful landings are more likely to land successfully in subsequent missions. This reinforces the value of reusability in space operations, not only from a cost perspective but also in terms of operational reliability. The analysis also highlighted the variation in success rates across different launch sites, with KSC LC 39A emerging as the most reliable site for successful recoveries. This insight could guide future decisions on launch site selection based on specific mission requirements.

5.1.3 Research Question 3: Enhancing Cost-Efficiency with Predictive Analytics

The predictive modelling for cost-efficiency showed that successful landings and increased reusability are strongly correlated with cost savings. The Random Forest Regressor achieved an R^2 score of 0.826, indicating that the model could explain a significant portion of the variance in the hypothetical cost savings variable. This finding supports the hypothesis that predictive analytics can enhance cost-efficiency strategies by accurately forecasting cost savings based on key operational factors. The scenario analysis further demonstrated that strategies focusing on increasing reusability and ensuring successful landings could lead to substantial financial benefits. These insights are crucial for companies like SpaceX, as they highlight the importance of integrating predictive analytics into their operational strategies to optimise costs and improve overall efficiency.

5.2 Biases and Limitations of the Study

5.2.1 Data-Related Biases

One potential bias in this study is the *sampling bias* due to the dataset being limited to SpaceX launches. While SpaceX is a major player in the space industry, the findings may not be generalisable to other companies or future launch conditions. Additionally, the dataset did not include variables such as weather conditions, which could significantly influence landing success. This *feature selection bias* means that the models may not account for all factors impacting launch outcomes, potentially limiting the robustness of the predictions *(Singhi and Liu, n.d.*).

5.2.2 Model-Related Limitations

The study utilised several machine learning models, each with its own strengths and limitations. The *Random Forest* model, while achieving high accuracy, is often criticised for being a "black box" due to its complexity and lack of interpretability *(Wikipedia Contributors, 2019).* This limitation may hinder the ability to understand the underlying decision-making process fully, making it challenging to apply these insights in practical scenarios. Similarly, the *Neural Network* model, although effective, requires a large amount of data to generalise well, which may not always be available in the context of space launches *(Wikipedia Contributors, 2018).*

5.2.3 Generalisability and Overfitting

Another limitation is the *generalisability* of the models. The models were trained and validated on a specific dataset, and their performance might vary when applied to new or unseen data, particularly if the conditions differ significantly from those captured in the dataset. The perfect accuracy achieved by the Random Forest model also raises concerns about *overfitting*—where the model performs exceptionally well on the training data but may not generalise to new data. This overfitting could reduce the model's utility in real-world applications where conditions are variable and less controlled *(Wikipedia Contributors, 2019).*

5.2.4 Data Availability and Size

The study was also constrained by the *limited dataset size*, with only 90 entries. While the models performed well, the small sample size may not capture the full variability and complexity of space launches. This limitation could affect the reliability of the predictions, particularly when extrapolating to future missions with different parameters *(Wikipedia Contributors, 2019).*

5.3 Contributions to Sustainable Development Goals (SDGs)

5.3.1 SDG 9: Industry, Innovation, and Infrastructure

This research contributes to *SDG 9* by providing actionable insights into improving the success rates of space launches through predictive modelling. By identifying key factors that influence landing success and cost-efficiency, the study supports the advancement of *technological innovation* in the aerospace industry. The use of machine learning models like Random Forest and XGBoost exemplifies how cutting-edge technology can be leveraged to enhance the reliability and efficiency of space operations, thereby strengthening industry infrastructure *(United Nations, 2015).*

5.3.2 SDG 13: Climate Action

The findings also have implications for *SDG 13*, which focuses on climate action. The study's emphasis on improving reusability directly relates to reducing the environmental impact of space launches. By increasing the frequency of successful landings and booster reusability, the research supports more sustainable space operations, reducing the need for new materials and lowering the carbon footprint associated with manufacturing and launching new rockets *(United Nations, 2015).*

5.3.3 SDG 17: Partnerships for the Goals

Finally, the research aligns with *SDG 17*, which emphasises partnerships to achieve global goals. The insights gained from this study could foster collaboration between private space companies and government agencies, promoting shared knowledge and joint efforts to improve space launch success rates and sustainability. Such partnerships are crucial for advancing space exploration in a way that is economically and environmentally sustainable *(United Nations, 2015).*

5.4 Recommendations for Future Research

5.4.1 Broader Data Collection

Future research should aim to expand the dataset to include a wider range of variables, such as weather conditions, technical specifications, and real-time operational data. Incorporating these additional factors could provide a more comprehensive understanding of the variables influencing landing success and cost-efficiency, thereby enhancing the robustness of the predictive models.

5.4.2 Advanced Modelling Techniques

While this study focused on Random Forest and XGBoost, future research could explore other *ensemble methods* and *deep learning models* to potentially improve predictive accuracy. Techniques such as *Gradient Boosting Machines (GBM)* or *ensemble methods combining multiple models* could offer better performance, particularly with larger and more diverse datasets *(Wikipedia Contributors, 2019).*

5.4.3 Integration of Real-Time Data

There is also potential for integrating real-time data feeds into the predictive models. This would enable dynamic forecasting and allow space companies to adjust their strategies in real-time based on current conditions, further optimising landing success rates and cost-efficiency.

While the study has limitations, including potential biases and a small dataset, its findings offer a solid foundation for future research. The contributions to SDGs underscore the broader impact of this research, advocating for continued innovation and collaboration in the aerospace industry to achieve sustainable and reliable space exploration.

6. Conclusion

This research aimed to explore and enhance the prediction of Falcon 9 first stage landings, understand the technical and operational factors influencing recovery success, and develop strategies for cost-efficiency through predictive analytics. The study leveraged advanced machine learning techniques, analysed critical features affecting landing outcomes, and examined how predictive models can drive cost savings in space launch operations.

6.1 Recap of Key Insights

The research confirmed that machine learning models, particularly Random Forest and XGBoost, are highly effective in predicting the successful landing of Falcon 9's first stage. These models outperformed traditional statistical approaches, with Random Forest achieving perfect accuracy in its predictions. The study identified critical features such as GridFins, ReusedCount, and LandingPad, which significantly influence landing outcomes *(see Appendix A)*. The strong correlation between these features and landing success underscores the importance of incorporating advanced analytics into space mission planning.

In addition to predictive modelling, the research highlighted the significance of operational and technical factors in the success of rocket recoveries. The presence of GridFins and Legs, along with a higher ReusedCount, were found to be crucial for improving recovery rates. The analysis of launch sites further revealed that KSC LC 39A had the highest success rate, emphasising the role of site-specific conditions in mission outcomes.

6.2 Implications and Applications

The findings of this study have important implications for the aerospace industry, particularly for companies like SpaceX that rely heavily on the reusability of rockets to drive down costs. By integrating predictive analytics into their operations, these companies can enhance the reliability and efficiency of their missions. The identification of key predictors of landing success allows for more informed decision-making in both pre-launch planning and post-launch analysis. Furthermore, the cost-efficiency analysis demonstrates the financial benefits of optimising reusability, providing a strong business case for the continued investment in predictive analytics *(Peter, 2016)*.

6.3 Managerial Recommendations

Based on the findings, several recommendations can be made for industry professionals. First, it is essential to prioritise the features identified as significant predictors in both the design and operational stages of rocket launches. For instance, ensuring the presence of GridFins and maximising the reuse of rockets should be key considerations in mission planning *(see Appendix A)*. Additionally, continuous data collection and model refinement are crucial for maintaining the accuracy and reliability of predictive models. As new data becomes available, it should be integrated into existing models to ensure they remain relevant and effective in predicting outcomes.

6.4 Future Research Directions

Building on this research, future studies could delve into the integration of real-time data streams, such as live telemetry and environmental conditions, to refine and enhance predictive models. Additionally, expanding the scope to include predictive analytics for other space mission components, like payload optimisation and interplanetary missions, could yield further insights. Investigating the scalability of these models for use in next-generation rockets and spacecraft will be crucial as the industry advances. Lastly, as machine learning evolves, future research should focus on developing adaptive models that can learn and improve from new data autonomously *(Hassanien, Darwish & Abdelghafar, 2019).*

In conclusion, this research has demonstrated the significant benefits of integrating predictive analytics into space operations. By accurately forecasting landing success and optimising cost-efficiency strategies, these advanced techniques offer a powerful tool for improving the reliability and sustainability of space missions. As the aerospace industry continues to evolve, the insights gained from this study will be invaluable in guiding future developments and ensuring the continued success of space exploration efforts.

7. Appendix A

Comprehensive Description of the SpaceX Falcon 9 Launch Dataset

The dataset under analysis contains data on 90 SpaceX Falcon 9 launches, with 18 variables that capture different facets of these missions. Below, each column in the dataset is thoroughly explained, covering every detail to ensure a comprehensive understanding.

A.1. Unnamed: 0

The Unnamed: 0 column is an automatically generated index column that sequentially numbers each row in the dataset. It starts at 0 and ends at 89, corresponding to the total number of entries (90).

A.2. FlightNumber

The FlightNumber column contains an integer value representing the sequential number assigned to each Falcon 9 launch. It ranges from 1 to 90, reflecting the chronological order in which these launches occurred.

A.3. Date

The Date column records the date on which each launch occurred, formatted as a string in the YYYY-MM-DD format (e.g., 2010-06-04 for 4th June 2010).

A.4. BoosterVersion

The BoosterVersion column specifies the version of the Falcon 9 rocket used in each launch. In this dataset, all entries indicate the use of the Falcon 9 booster, a key member of SpaceX's family of rockets.

A.5. PayloadMass

The PayloadMass column represents the mass of the payload delivered to orbit by the Falcon 9 rocket, measured in kilograms. This variable is critical for evaluating the rocket's performance, as the payload mass directly influences the energy required for the launch and the trajectory to the target orbit. The dataset records payload masses ranging from 350 kg to 15,600 kg, with an average payload mass of approximately 6,123.5 kg. However, it is important to note that 5 entries have missing values in this column, which may be due to the absence of recorded data for those launches or missions where payload mass was not disclosed.

A.6. Orbit

The Orbit column indicates the type of orbit the payload was intended to reach. This column includes a variety of orbital classifications, such as:

- **LEO (Low Earth Orbit)**: A commonly used orbit located relatively close to Earth, typically at altitudes of 200 to 2,000 kilometres. LEO is often used for Earth observation satellites, the International Space Station (ISS), and some scientific missions.

- **GTO (Geostationary Transfer Orbit)**: An intermediate orbit used to transfer payloads to a geostationary orbit. GTO is a highly elliptical orbit that enables the payload to reach the geostationary belt, where it can remain fixed relative to a point on Earth.
- **ISS (International Space Station)**: A specific LEO orbit where the payload is intended to reach or dock with the ISS.
- **PO (Polar Orbit)**: An orbit that passes over the Earth's poles, enabling global coverage of the planet's surface. Polar orbits are often used for environmental monitoring and reconnaissance.

The type of orbit affects the mission's complexity, energy requirements, and the overall strategy for the launch

A.7. LaunchSite

The LaunchSite column provides the specific location from which each Falcon 9 rocket was launched. The dataset includes several launch sites, primarily in the United States, each with its own strategic importance:

- **CCSFS SLC 40 (Cape Canaveral Space Force Station Space Launch Complex 40)**: Located in Florida, this site is one of the most frequently used by SpaceX. Its proximity to the equator allows rockets to take advantage of the Earth's rotational speed, making it ideal for launches to low-inclination orbits such as LEO and GTO.
- **VAFB SLC 4E (Vandenberg Air Force Base Space Launch Complex 4E)**: Located in California, this site is used primarily for launches to polar orbits. Its location on the west coast allows rockets to launch southward over the Pacific Ocean, avoiding populated areas during ascent.
- **KSC LC 39A (Kennedy Space Center Launch Complex 39A)**: Also in Florida, this historic site was originally built for the Apollo missions and has been modified for use by SpaceX. It is used for a variety of mission types, including crewed flights to the ISS.

The choice of launch site is influenced by the mission's target orbit, with each site offering specific advantages in terms of trajectory and energy efficiency.

A.8. Outcome

The Outcome column documents the result of each launch and the subsequent landing attempt, if applicable. It provides a detailed account of both the mission success and the recovery effort for the Falcon 9 booster:

- **None None**: Indicates that no landing was attempted. This was common in early Falcon 9 missions where the focus was solely on delivering the payload to orbit, with no effort made to recover the booster.
- **True Ocean**: Signifies a successful landing attempt where the booster was intended to land but instead ended up in the ocean.
- **True ASDS (Autonomous Spaceport Drone Ship)**: Indicates a successful landing of the booster on a drone ship stationed in the ocean.
- **True RTLS (Return to Launch Site)**: Reflects a successful landing of the booster back at the launch site.

The Outcome column is crucial for analysing the reliability and success rate of SpaceX's missions, as well as the progression of its reusability efforts.

A.9. Flights

The Flights column indicates the number of times a specific Falcon 9 booster has been flown. This integer value highlights the reusability aspect of the Falcon 9, a key feature of SpaceX's operational strategy. In the dataset, this column shows that most boosters have been flown between 1 and 2 times, with some having been reused up to 6 times.

A.10. GridFins

The GridFins column is a Boolean variable (True or False) that indicates whether grid fins were used during the launch. Grid fins are aerodynamic control surfaces deployed during the booster's descent, enabling precise manoeuvring to guide the booster towards its landing site. These fins are crucial for missions where the booster is intended to be recovered, particularly when landing on a drone ship or at the launch site.

A.11. Reused

The Reused column is a Boolean variable (True or False) that indicates whether the booster used in the launch was reused from a previous mission. A True value in this column signifies that the booster had previously flown and was successfully refurbished for another mission, demonstrating the viability of SpaceX's approach to reducing the cost of spaceflight.

A.12. Legs

The Legs column is a Boolean variable (True or False) indicating whether the landing legs were deployed during the booster's descent. Landing legs are crucial for enabling vertical landings on solid ground or a drone ship, making them essential for missions where booster recovery is intended.

A.13. LandingPad

The LandingPad column indicates the designated site where the booster was intended to land, identified by a specific pad name. A value in this column suggests a planned recovery on a solid surface or drone ship, while a missing value often indicates no landing attempt or an ocean landing.

A.14. Block

The Block column indicates the block number of the Falcon 9 booster, which represents the specific version or iteration of the rocket. SpaceX's Falcon 9 rocket has undergone several block upgrades, with each block reflecting improvements in design, performance, and reusability. The block numbers in this dataset range from 1 to 5:

- **Block 1**: The original version of the Falcon 9, used in early missions. It had limited reusability and lacked the advanced features present in later blocks.
- **Block 2 and Block 3**: These versions introduced incremental improvements in engine performance and structural design, allowing for more efficient launches.
- **Block 4**: This block featured significant upgrades aimed at improving reusability, including stronger landing legs and a more robust thermal protection system.

- **Block 5**: The latest and most advanced version in the dataset, Block 5 was designed for extensive reusability, with features such as enhanced thermal protection, more durable engines, and streamlined refurbishment processes.

A.15. ReusedCount

The ReusedCount column captures the number of times a particular Falcon 9 booster has been reused. This integer value is a direct indicator of the booster's reusability and reflects the extent to which SpaceX has successfully implemented its strategy of reusing rocket hardware to reduce costs. The values in this column range from 0 (indicating a brand-new booster) to 12 (indicating a booster that has been reused multiple times).

- **0**: A value of 0 indicates that the booster was used for the first time in that launch.
- **1-12**: These values indicate the number of times the booster has been flown before the current mission. A higher reused count suggests a booster that has undergone multiple flights and refurbishments.

A.16. Serial

The Serial column provides the unique serial number assigned to each Falcon 9 booster. This serial number is a unique identifier that allows SpaceX and analysts to track the history and performance of individual boosters across multiple flights.

A.17. Longitude

The Longitude column records the geographic longitude coordinate of the launch site. Longitude values in this dataset range from -120.610829 to -80.577366 degrees, corresponding to the different launch sites used by SpaceX:

- **CCSFS SLC 40 and KSC LC 39A**: These sites are in Florida, with longitudes around -80.577366 degrees.
- **VAFB SLC 4E**: Located in California, this site has a longitude of approximately -120.610829 degrees.

A.18. Latitude

The Latitude column records the geographic latitude coordinate of the launch site. Latitude values in this dataset range from 28.561857 to 34.632093 degrees:

- **CCSFS SLC 40 and KSC LC 39A**: Located in Florida, these sites have latitudes around 28.561857 degrees.
- **VAFB SLC 4E**: Located in California, this site has a latitude of approximately 34.632093 degrees.

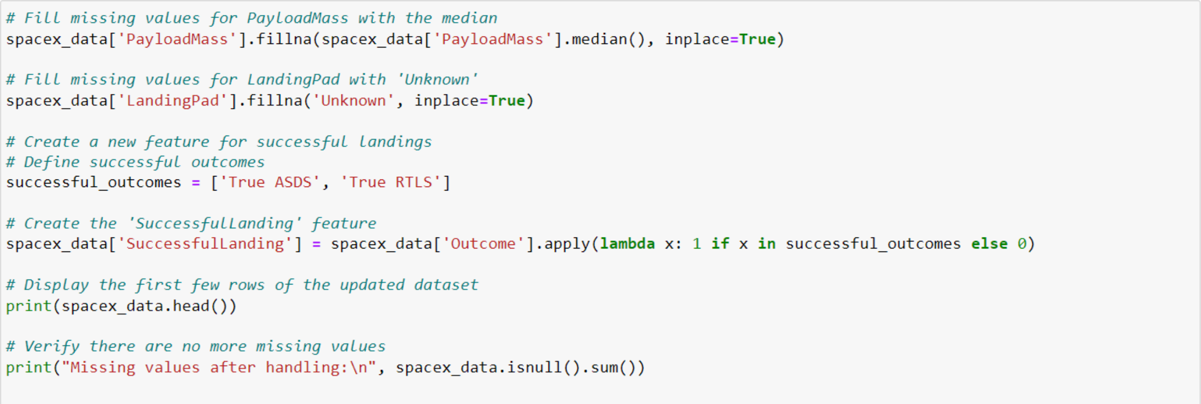
A.19. Summary Statistics and Data Quality

The dataset is largely complete, with most variables fully populated across all entries. However, there are some missing values in the PayloadMass and LandingPad columns. These missing values could be due to specific missions where the payload mass was not recorded or where the booster did not land on a designated pad.

8. Appendix B

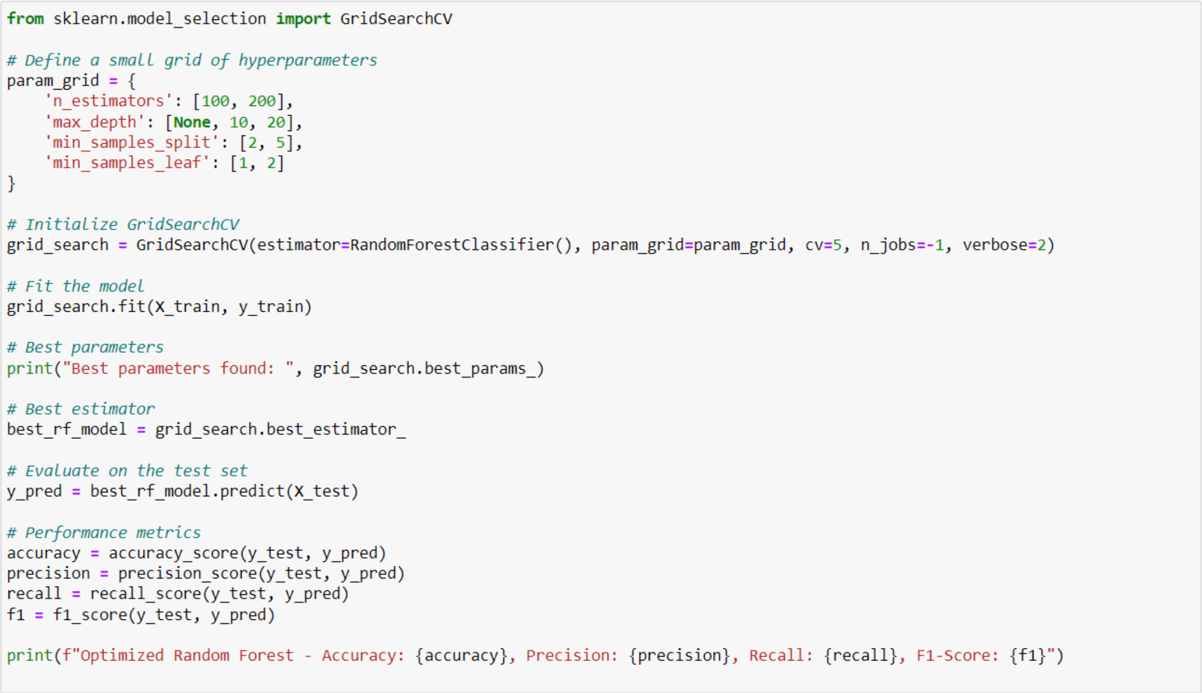
Key Code Implementations for Predictive Modelling and Analysis

B.1. Code: Data Preprocessing and Feature Engineering | Handling Missing Values and Feature Engineering



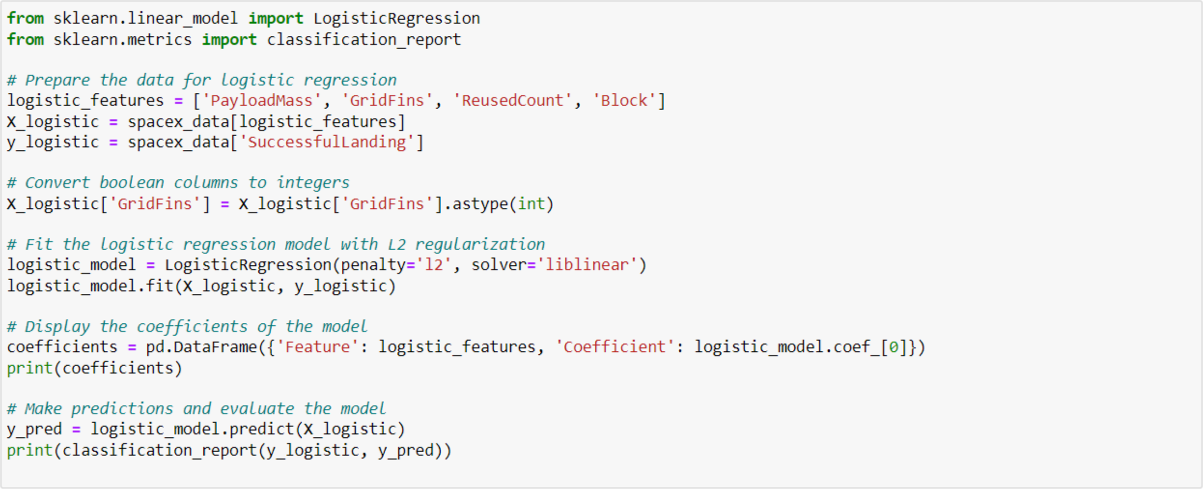
```
# Fill missing values for PayloadMass with the median
spacex_data['PayloadMass'].fillna(spacex_data['PayloadMass'].median(), inplace=True)

# Fill missing values for LandingPad with 'Unknown'
spacex_data['LandingPad'].fillna('Unknown', inplace=True)

# Create a new feature for successful landings
# Define successful outcomes
successful_outcomes = ['True ASDS', 'True RTLS']

# Create the 'SuccessfulLanding' feature
spacex_data['SuccessfulLanding'] = spacex_data['Outcome'].apply(lambda x: 1 if x in successful_outcomes else 0)

# Display the first few rows of the updated dataset
print(spacex_data.head())

# Verify there are no more missing values
print("Missing values after handling:\n", spacex_data.isnull().sum())
```

B.2. Code: Random Forest Model Training, Hyperparameter Tuning



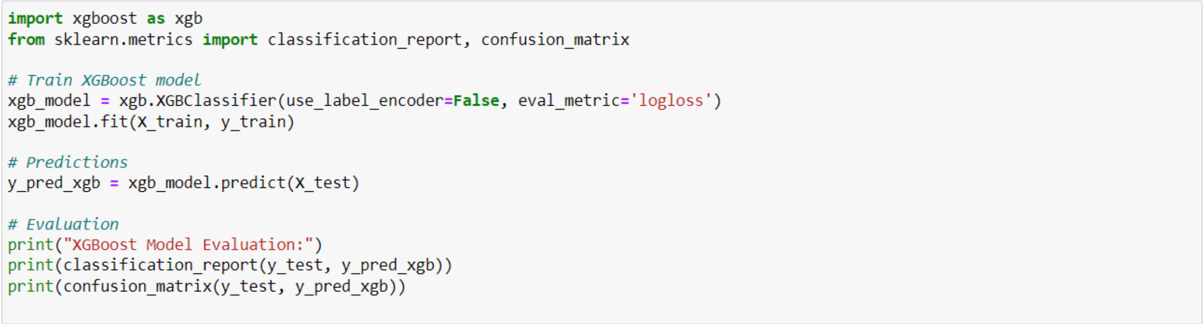
```
from sklearn.model_selection import GridSearchCV

# Define a small grid of hyperparameters
param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5],
    'min_samples_leaf': [1, 2]
}

# Initialize GridSearchCV
grid_search = GridSearchCV(estimator=RandomForestClassifier(), param_grid=param_grid, cv=5, n_jobs=-1, verbose=2)

# Fit the model
grid_search.fit(X_train, y_train)

# Best parameters
print("Best parameters found: ", grid_search.best_params_)

# Best estimator
best_rf_model = grid_search.best_estimator_

# Evaluate on the test set
y_pred = best_rf_model.predict(X_test)

# Performance metrics
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

print(f"Optimized Random Forest - Accuracy: {accuracy}, Precision: {precision}, Recall: {recall}, F1-Score: {f1}")
```

B.3. Code: Logistic Regression Model Training and Evaluation



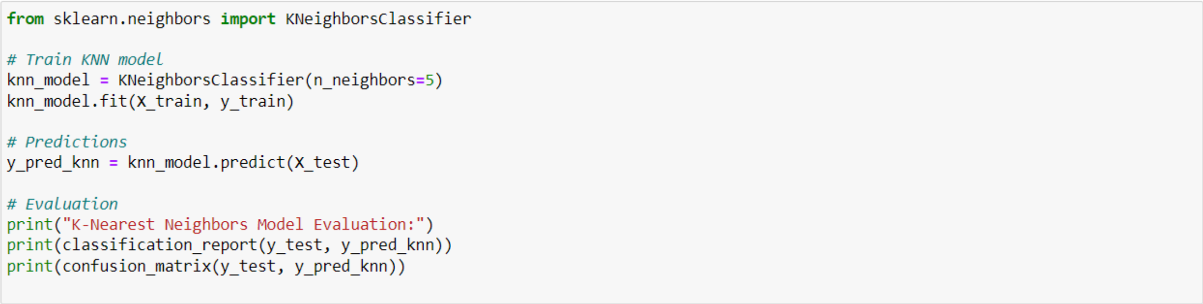
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report

# Prepare the data for logistic regression
logistic_features = ['PayloadMass', 'GridFins', 'ReusedCount', 'Block']
X_logistic = spacex_data[logistic_features]
y_logistic = spacex_data['SuccessfulLanding']

# Convert boolean columns to integers
X_logistic['GridFins'] = X_logistic['GridFins'].astype(int)

# Fit the logistic regression model with L2 regularization
logistic_model = LogisticRegression(penalty='l2', solver='liblinear')
logistic_model.fit(X_logistic, y_logistic)

# Display the coefficients of the model
coefficients = pd.DataFrame({'Feature': logistic_features, 'Coefficient': logistic_model.coef_[0]})
print(coefficients)

# Make predictions and evaluate the model
y_pred = logistic_model.predict(X_logistic)
print(classification_report(y_logistic, y_pred))
```

B.4. Code: XGBoost Model Training and Evaluation



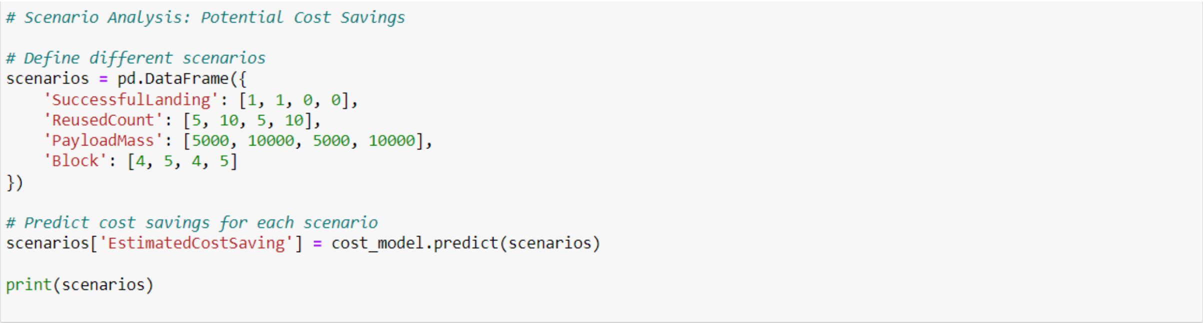
```
import xgboost as xgb
from sklearn.metrics import classification_report, confusion_matrix

# Train XGBoost model
xgb_model = xgb.XGBClassifier(use_label_encoder=False, eval_metric='logloss')
xgb_model.fit(X_train, y_train)

# Predictions
y_pred_xgb = xgb_model.predict(X_test)

# Evaluation
print("XGBoost Model Evaluation:")
print(classification_report(y_test, y_pred_xgb))
print(confusion_matrix(y_test, y_pred_xgb))
```

B.5. Code: K-Nearest Neighbors (KNN) Model Training and Evaluation



```
from sklearn.neighbors import KNeighborsClassifier

# Train KNN model
knn_model = KNeighborsClassifier(n_neighbors=5)
knn_model.fit(X_train, y_train)

# Predictions
y_pred_knn = knn_model.predict(X_test)

# Evaluation
print("K-Nearest Neighbors Model Evaluation:")
print(classification_report(y_test, y_pred_knn))
print(confusion_matrix(y_test, y_pred_knn))
```

B.6. Code: Scenario Analysis for Cost-Efficiency

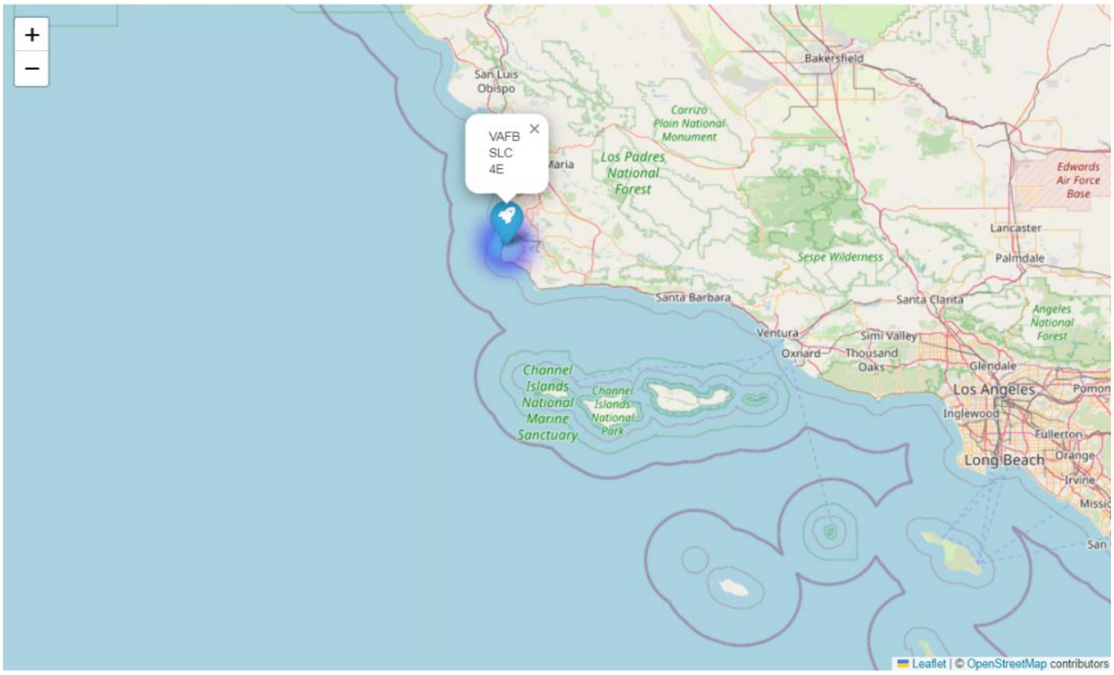


```
# Scenario Analysis: Potential Cost Savings

# Define different scenarios
scenarios = pd.DataFrame({
    'SuccessfulLanding': [1, 1, 0, 0],
    'ReusedCount': [5, 10, 5, 10],
    'PayloadMass': [5000, 10000, 5000, 10000],
    'Block': [4, 5, 4, 5]
})

# Predict cost savings for each scenario
scenarios['EstimatedCostSaving'] = cost_model.predict(scenarios)

print(scenarios)
```

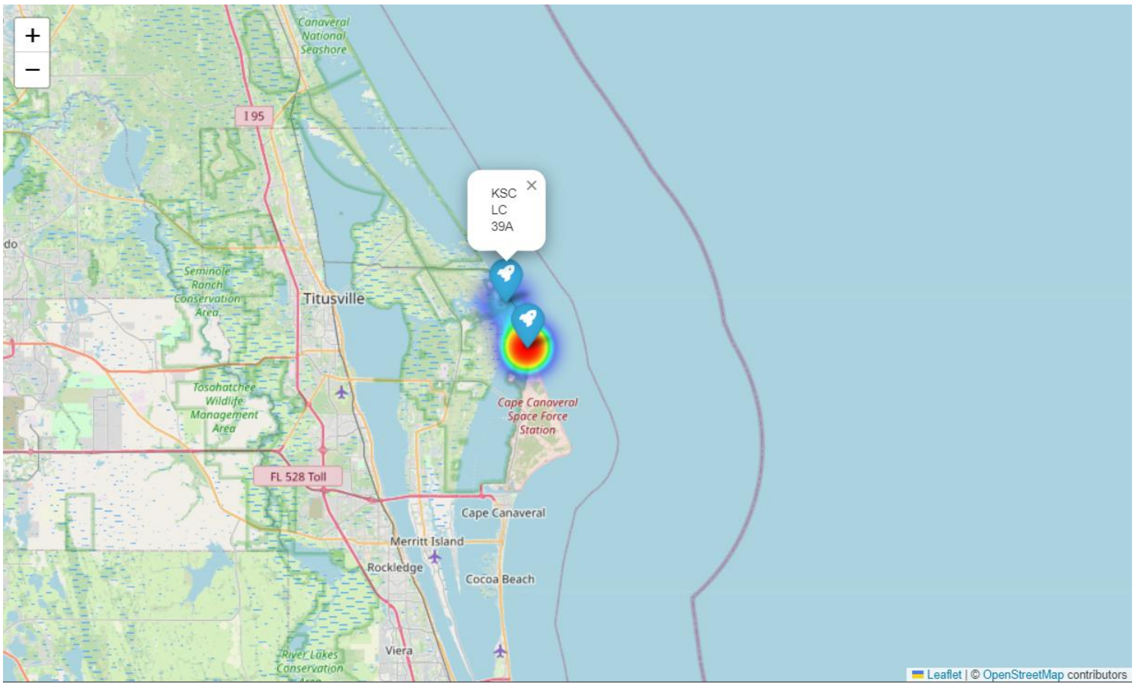
9. Appendix C

Folium Heatmap Launch Site Analysis

C.1. VAFB SLC 4E located on the west coast of California.

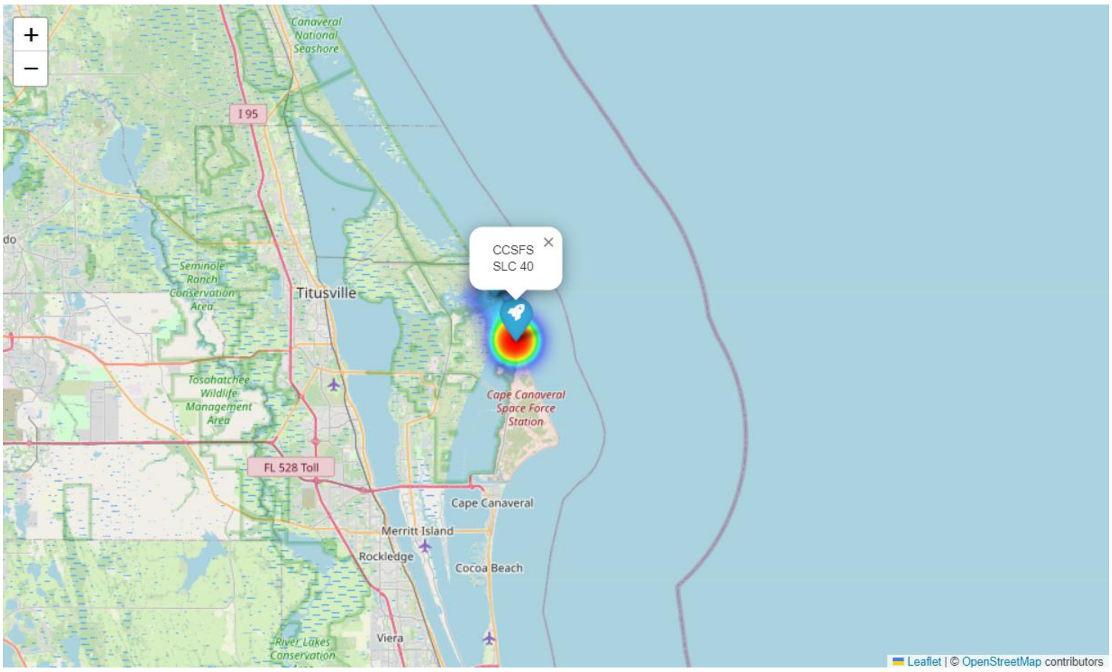


C.2. KSC LC 39A on the east coast of Florida.



KSC LC 39A
Canaveral National Seashore
I 95
Seminole Ranch Conservation Area
Titusville
Tosohatchee Wildlife Management Area
Cape Canaveral Space Force Station
FL 528 Toll
Cape Canaveral
Merritt Island
Rockledge
Cocoa Beach
Viera
River Lakes Conservation Area
Leaflet | © OpenStreetMap contributors

C.3. CCSFS SLC 40 located on the east coast of Florida.


+
−
Canaveral National Seashore
I 95
do
Seminole Ranch Conservation Area
Titusville
CCSFS SLC 40
×
Tosohatchee Wildlife Management Area
Cape Canaveral Space Force Station
FL 528 Toll
Cape Canaveral
Merritt Island
Rockledge
Cocoa Beach
River Lakes Conservation
Viera
Leaflet | © OpenStreetMap contributors

8. References

- Brown, E. (2023). *Boosting rocket reliability at the material level.* [online] MIT News | Massachusetts Institute of Technology. Available at: https://news.mit.edu/2023/boosting-rocket-reliability-material-level [Accessed 11 Jun. 2024].
- www.thespacereview.com. (n.d.). *The Space Review: The Falcon 9 achieves the shuttle's dreams*. [online] Available at: https://www.thespacereview.com/article/4542/1 [Accessed 12 Jun. 2024].
- Arnold, S. (2023). *How to Invest in Space in 2024*. [online] Private Equity Investing | Linqto Private Investing. Available at: https://www.linqto.com/blog/how-to-invest-in-space/ [Accessed 13 Jun. 2024].
- Aditya Singh Tharran (2023). *Introduction The aerospace industry is a complex and highly dynamic sector where safety and reliability are of paramount importance. In recent years, data science has emerged as a critical tool in enhancing the performance, safety, and efficiency of aircraft.* [online] Linkedin.com. Available at: https://www.linkedin.com/pulse/data-science-predictive-maintenance-aerospace-aditya-singh-tharran-zvyce [Accessed 15 Jun. 2024].
- Review, H.B. (2024). *The Economics of SpaceX*. [online] Hivelr. Available at: https://www.hivelr.com/2024/06/the-economics-of-spacex/ [Accessed 17 Jun. 2024].
- Jo, B.-U. and Ahn, J., 2021. Optimal staging of reusable launch vehicles considering velocity losses. *Aerospace Science and Technology*, 109, p.106431. https://doi.org/10.1016/j.ast.2020.106431 [Accessed 20 Jun. 2024].
- Tománek, R. and Hospodka, J., 2018. Reusable Launch Space Systems. *MAD - Magazine of Aviation Development*, 6, pp.10-13. Available at: https://www.researchgate.net/publication/330207086_Reusable_Launch_Space_Systems [Accessed 21 Jun. 2024].
- Prous, G. (n.d.). *Guidance and Control for Launch and Vertical Descend of Reusable Launchers using Model Predictive Control and Convex Optimisation*. [online] Available at: http://www.diva-portal.org/smash/get/diva2:1502202/FULLTEXT01.pdf [Accessed 25 Jun. 2024].
- SpaceX (2024). *Falcon 9*. [online] SpaceX. Available at: https://www.spacex.com/vehicles/falcon-9/ [Accessed 25 Jun. 2024].
- Lionnet, P. (2021). *Pierre Lionnet on LinkedIn: SpaceX launch services economics*. [online] Linkedin.com. Available at: https://www.linkedin.com/posts/eurospacepierrelionnet_spacex-launch-services-economics-activity-6840929023577612289-hzle/ [Accessed 27 Jun. 2024].
- Baiocco, P. (2021). Overview of reusable space systems with a look to technology aspects. *Acta Astronautica*, 189, pp.10–25. Available at: doi: https://doi.org/10.1016/j.actaastro.2021.07.039 [Accessed 1 Jul. 2024].
- Boiani, D., 2021. Main Structural Design Considerations for Reusable Launch Vehicles. *Webthesis*. Available at: https://webthesis.biblio.polito.it/18317/1/tesi.pdf [Accessed 2 Jul. 2024].
- Kaur, J. (2023). *Demystifying Machine Learning Algorithms: A Beginner's Guide*. [online] Xenonstack.com. Available at: https://www.xenonstack.com/blog/demystifying-machine-learning-algorithms [Accessed 3 Jul. 2024].
- Sorini, A., Pineda, E., Stuckner, J. and Gustafson, P. (n.d.). *A Convolutional Neural Network for Multiscale Modeling of Composite Materials*. [online] Available at:

https://ntrs.nasa.gov/api/citations/20205011437/downloads/202050011437%20Paper%20Final.pdf [Accessed 3 Jul. 2024].

- Sirohhi, A. (2024) "Machine Learning in Predictive Analytics". [online] Linkedin.com. Available at: https://www.linkedin.com/pulse/machine-learning-predictive-analytics-anjoum-sirohhi-av4tf [Accessed 10 Jul. 2024].
- Dataheadhunters.com. (2024). *How to build a predictive maintenance system in Python for aviation*. [online] Available at: https://dataheadhunters.com/academy/how-to-build-a-predictive-maintenance-system-in-python-for-aviation/ [Accessed 12 Jul. 2024].
- Verma, V., 2024. *Aircraft Predictive Maintenance: An Application of Machine Learning Algorithms*. Interim Report, July 2024. Available at: https://www.researchgate.net/publication/381924182_AIRCRAFT_PREDICTIVE_MAINTENANCE_AN_APPLICATION_OF_MACHINE_LEARNING_ALGORITHMS_VERSHA_INTERIM_REPORT_JULY_2024_2 [Accessed 13 Jul. 2024].
- Alfarhood, M., Alotaibi, R., Abdulrahim, B., Einieh, A., Almousa, M. and Alkhanifer, A. (2024). Predicting Flight Delays with Machine Learning: A Case Study from Saudi Arabian Airlines. *International Journal of Aerospace Engineering*, [online] 2024, p.e3385463. Available at: https://onlinelibrary.wiley.com/doi/10.1155/2024/3385463 [Accessed 14 Jul. 2024].
- Stanton, I., Munir, K., Ikram, A. and El-Bakry, M. (2022). Predictive maintenance analytics and implementation for aircraft: Challenges and opportunities. *Systems Engineering*, [online] 26(2). Available at: https://incose.onlinelibrary.wiley.com/doi/full/10.1002/sys.21651 [Accessed 14 Jul. 2024].
- Le Clainche, S., Ferrer, E., Gibson, S., Cross, E., Parente, A. and Vinuesa, R. (2023). Improving aircraft performance using machine learning: A review. *Aerospace Science and Technology*, [online] 138, p.108354. Available at: https://www.sciencedirect.com/science/article/pii/S1270963823002511 [Accessed 15 Jul. 2024].
- Pang, Y., Liu, Y., Yan, H., Zhuang, H., Marvi, H. and Ren, Y. (2023). *Artificial Intelligence-enhanced Predictive Modeling in Air Traffic Management*. [online] Available at: https://keep.lib.asu.edu/system/files/c7/Pang_asu_0010E_22693.pdf [Accessed 17 Jul. 2024].
- Muilwijk, T. (2024). *Predictive Maintenance: Why It Is Important to the Aerospace Industry*. [online] Royale International. Available at: https://www.royaleinternational.com/2024/02/predictive-maintenance-why-it-is-important-to-the-aerospace-industry/ [Accessed 18 Jul. 2024].
- Dupont, M. (2024). *Predictive Maintenance Revolutionized by AI Models*. [online] Labelvisor. Available at: https://www.labelvisor.com/predictive-maintenance-revolutionized-by-ai-models/ [Accessed 19 Jul. 2024].
- Kaplan, S. (2024). *Navigating the Future: The Role of Data Analytics in Aviation Manufacturing*. [online] Praxie.com. Available at: https://praxie.com/data-analytics-in-aviation-manufacturing/ [Accessed 19 Jul. 2024].
- Anon, (2024). *Smooth Sailing in the Sky: The Role of Predictive Maintenance in Aviation*. [online] Available at: https://praxie.com/predictive-maintenance-in-aviation/ [Accessed 20 Jul. 2024].
- Teubert, C., Pohya, A. and Gorospe, G. (2023). *An Analysis of Barriers Preventing the Widespread Adoption of Predictive and Prescriptive Maintenance in Aviation*. [online] Available at: https://ntrs.nasa.gov/api/citations/20230000841/downloads/Teubert_An%20Analysis%20of%20Barriers%20Preventing%20the%20Widespread%20Adoption%20of%20Predictive%20and%20Prescriptive%20Maintenance%20in%20Aviation%20(1).pdf [Accessed 20 Jul. 2024].

- Korba, P., Šváb, P., Vereš, M. and Lukáč, J. (2023). Optimizing Aviation Maintenance through Algorithmic Approach of Real-Life Data. *Applied Sciences*, [online] 13(6), p.3824. Available at: https://doi.org/10.3390/app13063824 [Accessed 20 Jul. 2024].
- BCG Global. (2020). *How Value Can Take Off with Predictive Aircraft Maintenance*. [online] Available at: https://www.bcg.com/publications/2020/building-value-with-predictive-aircraft-maintenance [Accessed 21 Jul. 2024].
- Suhir, E. (2014). Human-in-the-Loop (HITL): Probabilistic Predictive Modeling (PPM) of an Aerospace Mission/Situation Outcome. *Aerospace*, 1(3), pp.101–136. Available at:https://doi.org/10.3390/aerospace1030101 [Accessed 23 Jul. 2024].
- Verhagen, W.J.C., Santos, B.F., Freeman, F., van Kessel, P., Zarouchas, D., Loutas, T., Yeun, R.C.K. and Heiets, I. (2023). Condition-Based Maintenance in Aviation: Challenges and Opportunities. *Aerospace*, [online] 10(9), p.762. Available at: https://doi.org/10.3390/aerospace10090762 [Accessed 24 Jul. 2024].
- Com, K. (2020). *The future of space Communal, commercial, contested.* KPMG International. [online] p.2030. Available at: https://assets.kpmg.com/content/dam/kpmg/au/pdf/2020/30-voices-on-2030-future-of-space.pdf [Accessed 25 Jul. 2024].
- Denis, G., Alary, D., Pasco, X., Pisot, N., Texier, D. and Toulza, S. (2020). From new space to big space: How commercial space dream is becoming a reality. *Acta Astronautica*, [online] 166, pp.431–443. Available at: https://doi.org/10.1016/j.actaastro.2019.08.031 [Accessed 26 Jul. 2024].
- Sagar Varandekar (2022). *SpaceX_Falcon9_Launch_Data*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/sagarvarandekar/spacex-falcon9-launch-data?resource=download [Accessed 27 Jul. 2024].
- Foust, J., 2017. SpaceX gaining substantial cost savings from reused Falcon 9. *SpaceNews*, 5 April. Available at: SpaceX gaining substantial cost savings from reused Falcon 9 - SpaceNews [Accessed 28 Jul. 2024].
- United Nations (2015). *Transforming our world: the 2030 Agenda for Sustainable Development.:. Sustainable Development Knowledge Platform*. [online] United Nations. Available at: https://sustainabledevelopment.un.org/post2015/transformingourworld [Accessed 29 Jul. 2024].
- Wikipedia Contributors (2024). *Falcon 9*. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Falcon_9 [Accessed 30 Jul. 2024].
- Scikit-learn (2019). *sklearn.preprocessing.OneHotEncoder — scikit-learn 0.22 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html [Accessed 31 Jul. 2024].
- scikit learn (2018). *1.4. Support Vector Machines — scikit-learn 0.20.3 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/svm.html [Accessed 31 Jul. 2024].
- IBM (2023). *What Are Neural Networks?* [online] www.ibm.com. Available at: https://www.ibm.com/topics/neural-networks [Accessed 31 Jul. 2024].
- NVIDIA Data Science Glossary. (n.d.). *What is XGBoost?* [online] Available at: https://www.nvidia.com/en-gb/glossary/xgboost/ [Accessed 31 Jul. 2024].
- scikit-learn (2022). *1.10. Decision Trees — scikit-learn 0.22 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/tree.html [Accessed 1 Aug. 2024].

- IBM (2023). *What is Random Forest? | IBM.* [online] www.ibm.com. Available at: https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly [Accessed 3 Aug. 2024].
- IBM (n.d.). *What is Logistic regression? | IBM.* [online] www.ibm.com. Available at: https://www.ibm.com/topics/logistic-regression#:~:text=Logistic%20regression%20estimates%20the%20probability [Accessed 3 Aug. 2024].
- Chandrikasai (2023). *Imputing missing values is another technique used to handle missing data in a dataset.* [online] Medium. Available at: https://medium.com/@chandrikasai9997/imputing-missing-values-is-another-technique-used-to-handle-missing-data-in-a-dataset-824957ce71b4 [Accessed 4 Aug. 2024].
- Wikipedia. (2020). *Falcon 9 first-stage landing tests.* [online] Available at: https://en.wikipedia.org/wiki/Falcon_9_first-stage_landing_tests [Accessed 4 Aug. 2024].
- IBM (2023). *What is the k-nearest neighbors algorithm? | IBM.* [online] www.ibm.com. Available at: https://www.ibm.com/topics/knn [Accessed 4 Aug. 2024].
- Singhi, S. and Liu, H. (n.d.). *Feature Subset Selection Bias for Classification Learning.* [online] Available at: https://www.public.asu.edu/~huanliu/papers/icml06.pdf [Accessed 4 Aug. 2024].
- Wikipedia Contributors (2019). *Random forest.* [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Random_forest [Accessed 5 Aug. 2024].
- Wikipedia Contributors (2018). *Artificial neural network*. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Artificial_neural_network [Accessed 5 Aug. 2024].
- Wikipedia Contributors (2019). *Overfitting.* [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Overfitting [Accessed 5 Aug. 2024].
- Wikipedia Contributors (2019). *Sample size determination.* [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Sample_size_determination [Accessed 5 Aug. 2024].
- Peter (2016). *SpaceX's reusable Falcon 9: What are the real cost savings for customers?* [online] SpaceNews. Available at: https://spacenews.com/spacexs-reusable-falcon-9-what-are-the-real-cost-savings/ [Accessed 6 Aug. 2024].
- Hassanien, A.E., Darwish, A., & Abdelghafar, S. (2020) 'Machine learning in telemetry data mining of space mission: basics, challenges, and future directions', *Artificial Intelligence Review*, 53, pp. 3201–3230. https://doi.org/10.1007/s10462-019-09760-1 [Accessed 7 Aug. 2024].