**Comprehensive Analysis and Predictive Modeling of SpaceX Falcon 9 Launch Data**

**1. Introduction**

SpaceX Falcon 9 launches have revolutionized space exploration, emphasizing the importance of reusable rocket technology to reduce costs and improve efficiency. The aim of this project is to analyze the historical launch data of SpaceX Falcon 9, uncover patterns, and predict launch outcomes. By combining data cleaning, visualization, and machine learning, this analysis provides actionable insights for operational optimization.

**2. Objectives**

1. **Data Preparation:** To clean and preprocess SpaceX launch data, ensuring it is ready for analysis.
2. **Exploration:** To visualize key patterns and trends in the data, such as payload distributions, launch site frequencies, and reusability impacts.
3. **Modeling:** To apply machine learning techniques to predict launch success or failure based on relevant features.
4. **Evaluation:** To compare the performance of different machine learning models and interpret their strengths and weaknesses.

**3. Data Understanding**

The dataset comprises historical records of SpaceX Falcon 9 launches, with the following key features:

* **Date:** Date of each launch.
* **BoosterVersion:** Specific rocket version used.
* **PayloadMass:** Payload weight in kilograms, which impacts launch success and orbit placement.
* **LaunchSite:** Geographic site of launch, reflecting logistical considerations.
* **Orbit:** The intended orbital destination of the payload (e.g., LEO, GTO).
* **Outcome:** Whether the mission was successful or not, crucial for measuring SpaceX’s progress.
* **GridFins, Legs, Reused:** Indicators of rocket reusability, a hallmark of SpaceX’s strategy to reduce costs.

**Initial Observations:**

* The dataset has a mix of numerical and categorical variables.
* Missing values exist, particularly in payload and certain categorical columns.
* Class imbalance is evident in the target variable, as some outcomes (e.g., "True ASDS") occur more frequently than others.

**4. Data Cleaning and Preprocessing**

**Step 1: Handling Missing Data**

* **Numerical Variables:** Missing values in PayloadMass were replaced with the mean, ensuring no bias in averages.
* **Categorical Variables:** Missing values in LaunchSite and Orbit were imputed with the mode (most frequent value), as these are likely reflective of standard practices.

**Step 2: Data Transformation**

* The Date column was converted to a datetime format. From this, new columns (Year, Month, Day) were created to enable time-based analysis.
* One-hot encoding was applied to categorical columns (BoosterVersion, LaunchSite, Orbit), converting them into numerical formats suitable for machine learning models.

**Step 3: Outlier Detection**

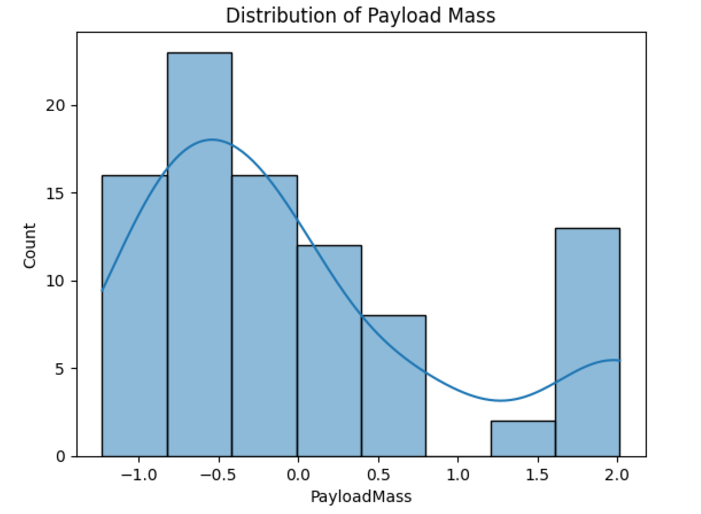
* Using Z-scores, outliers in PayloadMass were identified but were within acceptable bounds, indicating no need for removal.

**Step 4: Feature Scaling**

* StandardScaler was applied to PayloadMass, Longitude, and Latitude to normalize their scales, improving model performance.

**5. Data Visualization**

**5.1 Payload Mass Distribution**

****

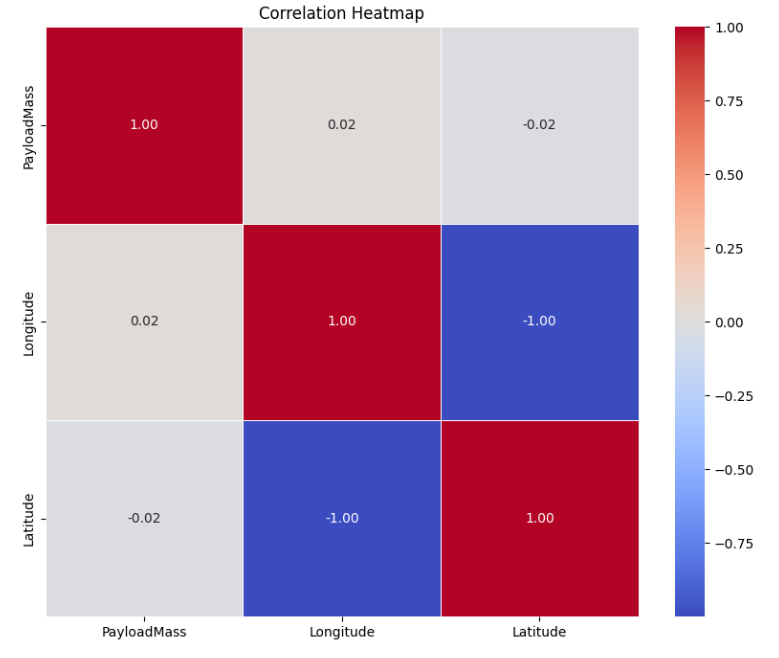
* A histogram revealed a skewed distribution of payload masses. Most payloads weighed between 500–3000 kg, with fewer instances of heavier payloads (>5000 kg). This trend reflects typical satellite and cargo sizes for various orbits.

**5.2 Launch Outcome Distribution**



* A count plot highlighted the dominance of successful launches (True ASDS) over failures (False Ocean, False RTLS). This illustrates SpaceX’s advancements in rocket reliability over time.

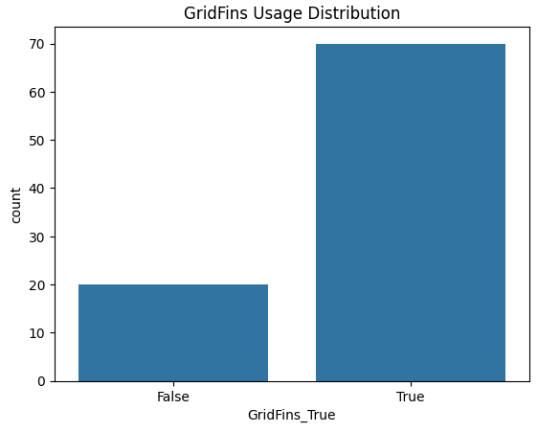
**5.3 Correlation Heatmap**

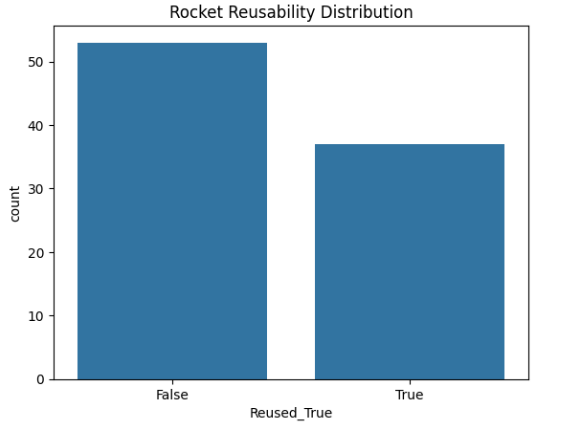


* A heatmap showed negligible correlation between PayloadMass, Longitude, and Latitude. This suggests that spatial variables and payload weight have minimal direct relationships but may interact with categorical variables like LaunchSite.

**5.4 Reusability Insights**

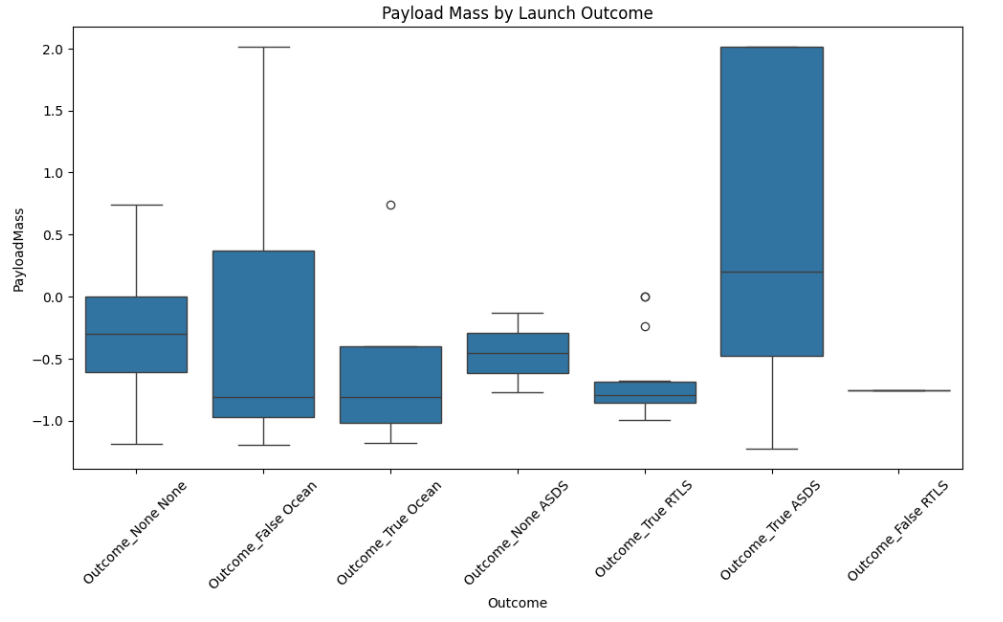
* Count plots for GridFins, Legs, and Reused revealed:





* + Rockets with grid fins and reusable components had significantly higher success rates.
  + Reusability metrics (Reused\_True) showed strong alignment with SpaceX’s goal of achieving cost-effective launches.

**5.5 Payload vs. Outcome**



* A boxplot comparing payload mass across outcomes showed:
  + Lighter payloads tended to succeed more frequently.
  + Heavier payloads (above 5000 kg) occasionally correlated with mission challenges, likely due to the complexity of achieving stable orbits for such payloads.

**6. Data Mining Techniques**

**Step 1: Defining Features and Target**

* Features (X) included preprocessed variables such as PayloadMass, Orbit, LaunchSite, and reusability metrics (GridFins, Reused).
* The target (y) represented the Outcome, specifically whether the launch was successful or not.

**Step 2: Handling Imbalanced Classes**

* SMOTE (Synthetic Minority Oversampling Technique) was applied to oversample rare classes, ensuring that models were not biased toward frequent outcomes.

**Step 3: Splitting Data**

* The data was split into training (80%) and testing (20%) subsets to evaluate model generalizability.

**Step 4: Machine Learning Models**

* **Logistic Regression:** Linear classification model that interprets relationships between features and outcomes.
* **Decision Tree Classifier:** Tree-based model for non-linear patterns and interactions.
* **Random Forest Classifier:** An ensemble method combining multiple decision trees for robustness.

**7. Results**

**Logistic Regression:**

* **Accuracy:** 66.67%
* **Strengths:** High interpretability and efficient computation.
* **Weaknesses:** Limited ability to model complex, non-linear relationships.

**Decision Tree Classifier:**

* **Accuracy:** 55.56%
* **Strengths:** Effective for non-linear data and feature interactions.
* **Weaknesses:** Prone to overfitting on small datasets.

**Random Forest Classifier:**

* **Accuracy:** 66.67%
* **Strengths:** Robust against overfitting and provides feature importance metrics.
* **Weaknesses:** Requires more computational resources.

**Classification Report:**

* All models struggled with rare outcomes (False Ocean, False RTLS) due to class imbalance. However, Random Forest achieved better recall and F1-scores for the majority class (True ASDS).

**8. Discussion**

The analysis highlights several key takeaways:

1. **Reusability Features:** Components like GridFins and Legs significantly improve mission success rates.
2. **Payload Mass Impact:** While payload mass generally does not correlate with outcomes, extreme payloads (either very light or very heavy) showed distinct success patterns.
3. **Model Selection:** Random Forest outperformed Logistic Regression and Decision Trees by handling complex feature interactions and reducing overfitting.

**Challenges:**

* Class imbalance limited the predictive power for rare outcomes.
* Missing variables such as weather or rocket-specific configurations reduced model accuracy.

**Future Directions:**

* Incorporate external factors (e.g., weather, mission complexity).
* Experiment with advanced algorithms like Gradient Boosting or Neural Networks for improved performance.

**9. Conclusion**

This report demonstrates the application of data mining and machine learning to analyze SpaceX Falcon 9 launch data. By cleaning, visualizing, and modeling the data, actionable insights were generated to optimize future launches. The findings emphasize the importance of reusable components and suggest areas for further investigation.

**10. References**

* **Libraries:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Imbalanced-learn.
* **Dataset:** SpaceX Falcon 9 Launch Data (publicly available).