Machine Learning Project 2

Jonathan Chacko Pattasseril

Regression and Classification Methods

Report

Table of Contents

[Introduction 1](#_Toc193679288)

[Regression Task: Flight Price Prediction 2](#_Toc193679289)

[Dataset Description and Objective 2](#_Toc193679290)

[Data Understanding 2](#_Toc193679291)

[Model Implementation & Evaluation 4](#_Toc193679292)

[Results Visualization 6](#_Toc193679293)

[Model Interpretation 7](#_Toc193679294)

[Classification Task: Phishing Websites Detection 8](#_Toc193679295)

[Dataset Description and Objective 8](#_Toc193679296)

[Data Understanding 8](#_Toc193679297)

[Model Implementation & Evaluation 10](#_Toc193679298)

[Results Visualization 12](#_Toc193679299)

[Model Interpretation 13](#_Toc193679300)

[Conclusion 14](#_Toc193679301)

[Summary of Findings 14](#_Toc193679302)

[Challenges and Limitations 14](#_Toc193679303)

[Real-world Implications 14](#_Toc193679304)

[Future Work 15](#_Toc193679305)

[References 15](#_Toc193679306)

# Introduction

This report presents the application of various machine learning techniques for regression and classification tasks. Two distinct datasets were selected for analysis:

1. **Flight Price Prediction dataset** (Regression task) - To predict airline ticket prices based on features like airline, flight details, source/destination cities, etc.
2. **Phishing Websites dataset** (Classification task) - To classify websites as legitimate or phishing based on URL and website features.

The analysis follows a structured approach including data exploration, preprocessing, model implementation, evaluation, and interpretation, providing insights into the performance of different machine learning algorithms for each task.

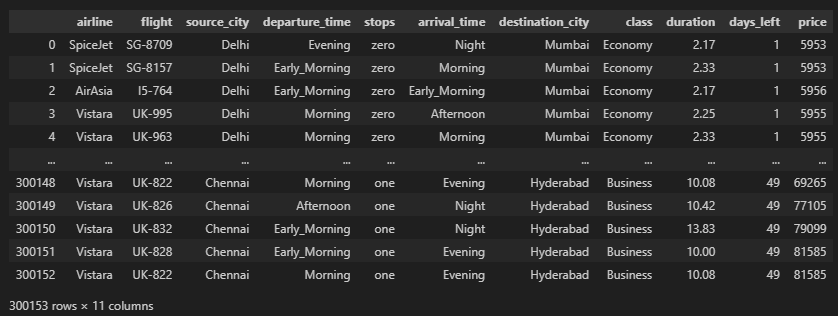
# Regression Task: Flight Price Prediction

## Dataset Description and Objective

**Source**: Kaggle - Flight Price Prediction dataset <https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction>

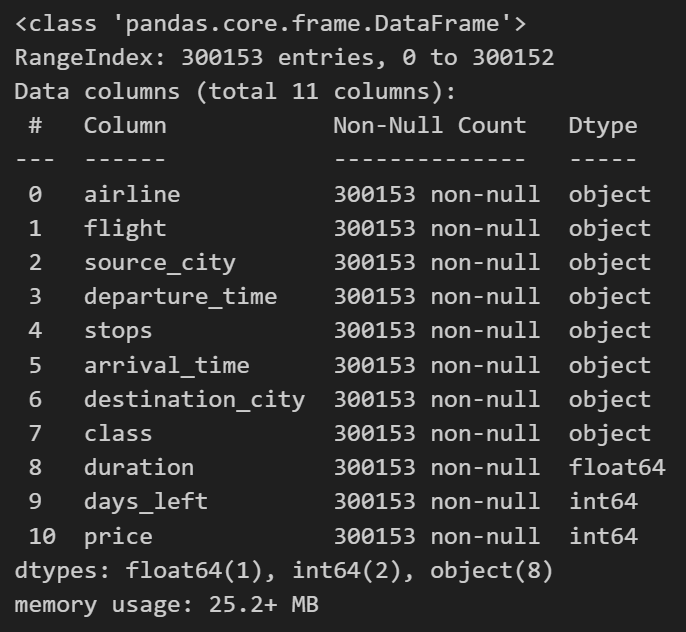
**Objective**: To predict the price of airline tickets based on various features such as airline, flight details, source and destination cities, departure and arrival times, number of stops, class, duration, and days left before the flight. This regression model aims to help travelers anticipate flight costs and potentially make more economical travel decisions.

## Data Understanding

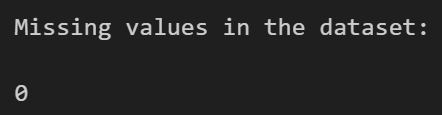


The dataset contains information about flight prices with multiple features that influence the cost of air travel. Initial exploration revealed:

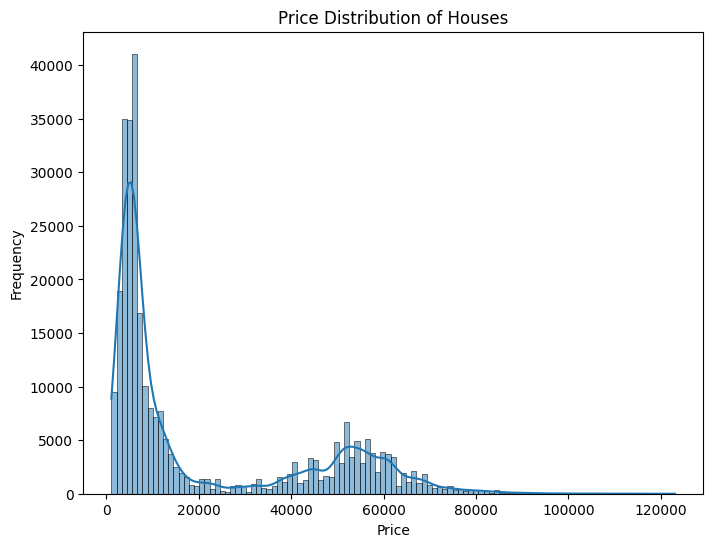
* Dataset shape: The dataset contains a substantial number of records with multiple features **(300153, 11)**

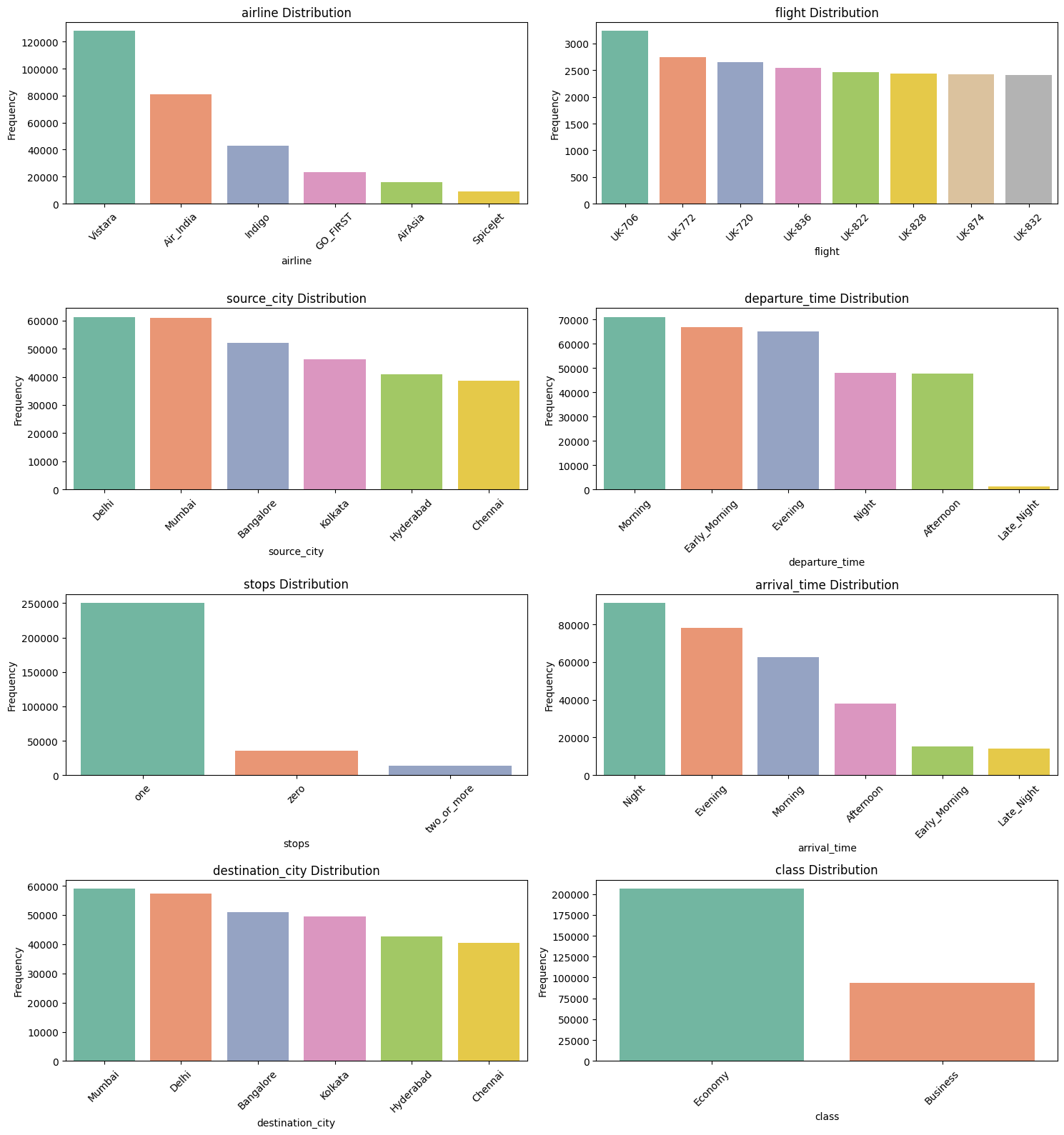


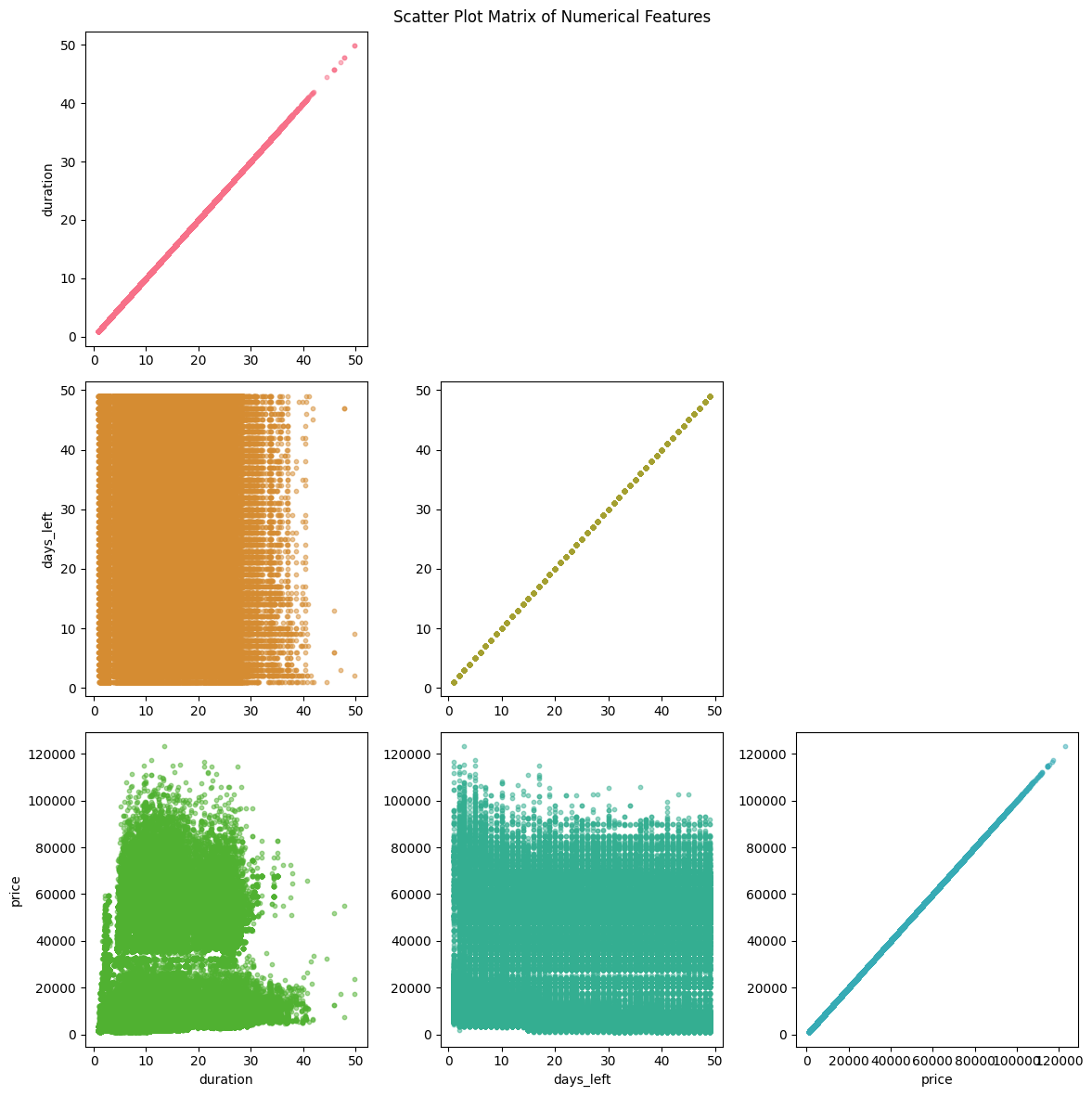
* No missing values were detected in the dataset



* The price distribution shows a right-skewed pattern, indicating most flights are in the lower price range with some expensive outliers



* The Object/Categorical Variables & their Data:
* Numeric Features Correlation to each other:



## Model Implementation & Evaluation

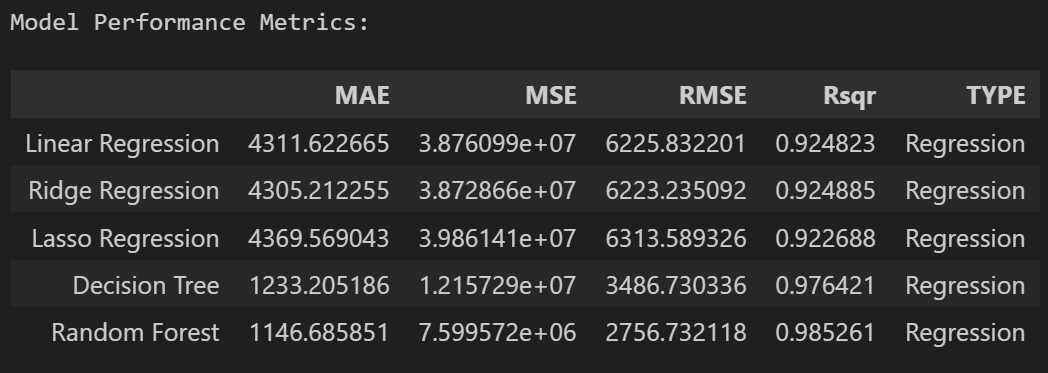
Five regression models were implemented and evaluated:

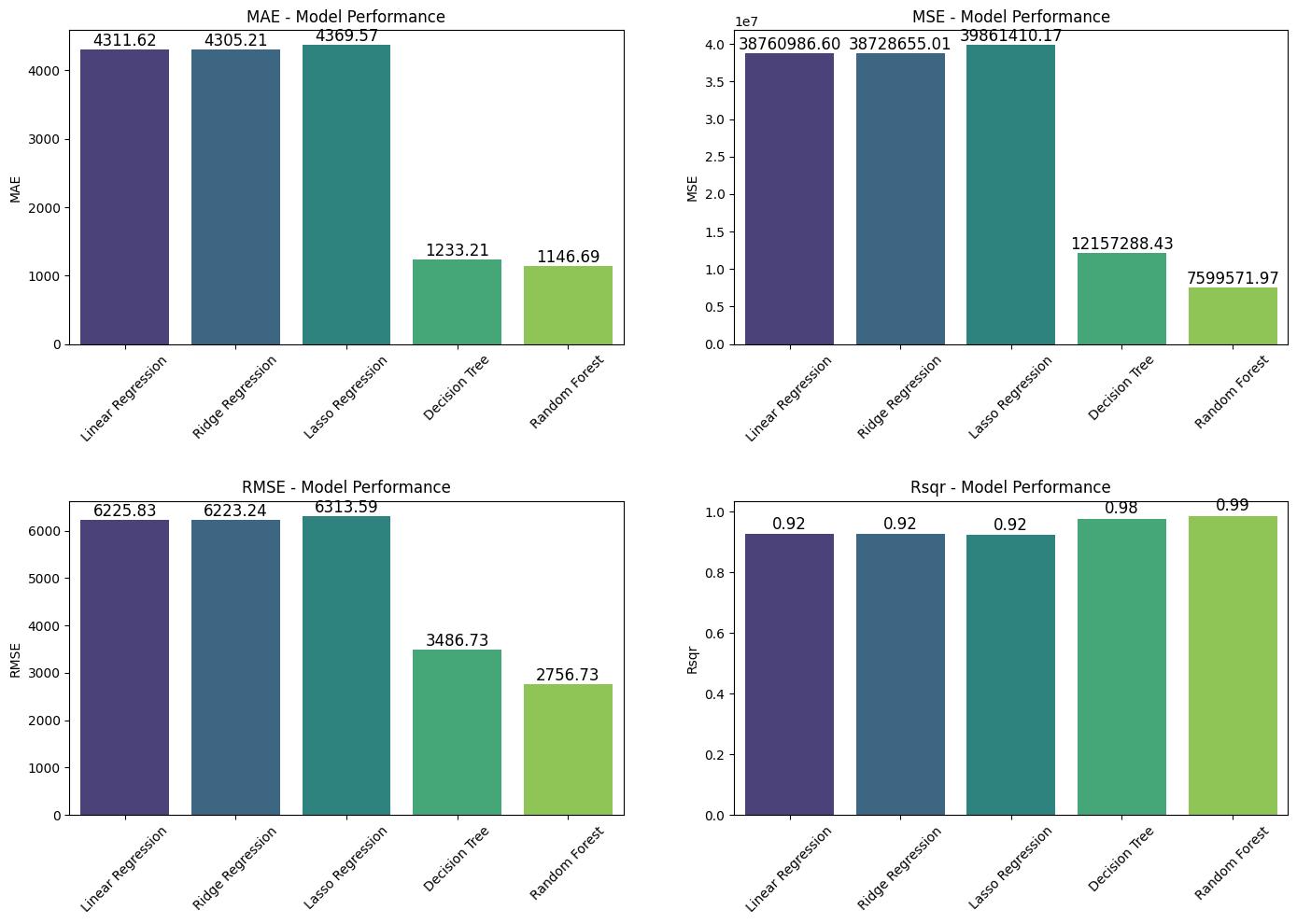
1. Linear Regression
2. Ridge Regression
3. Lasso Regression
4. Decision Tree Regressor
5. Random Forest Regressor

The models were evaluated using the following metrics:

* Mean Absolute Error (MAE)
* Mean Squared Error (MSE)
* Root Mean Squared Error (RMSE)
* R-squared (R²)

Results of model performance:



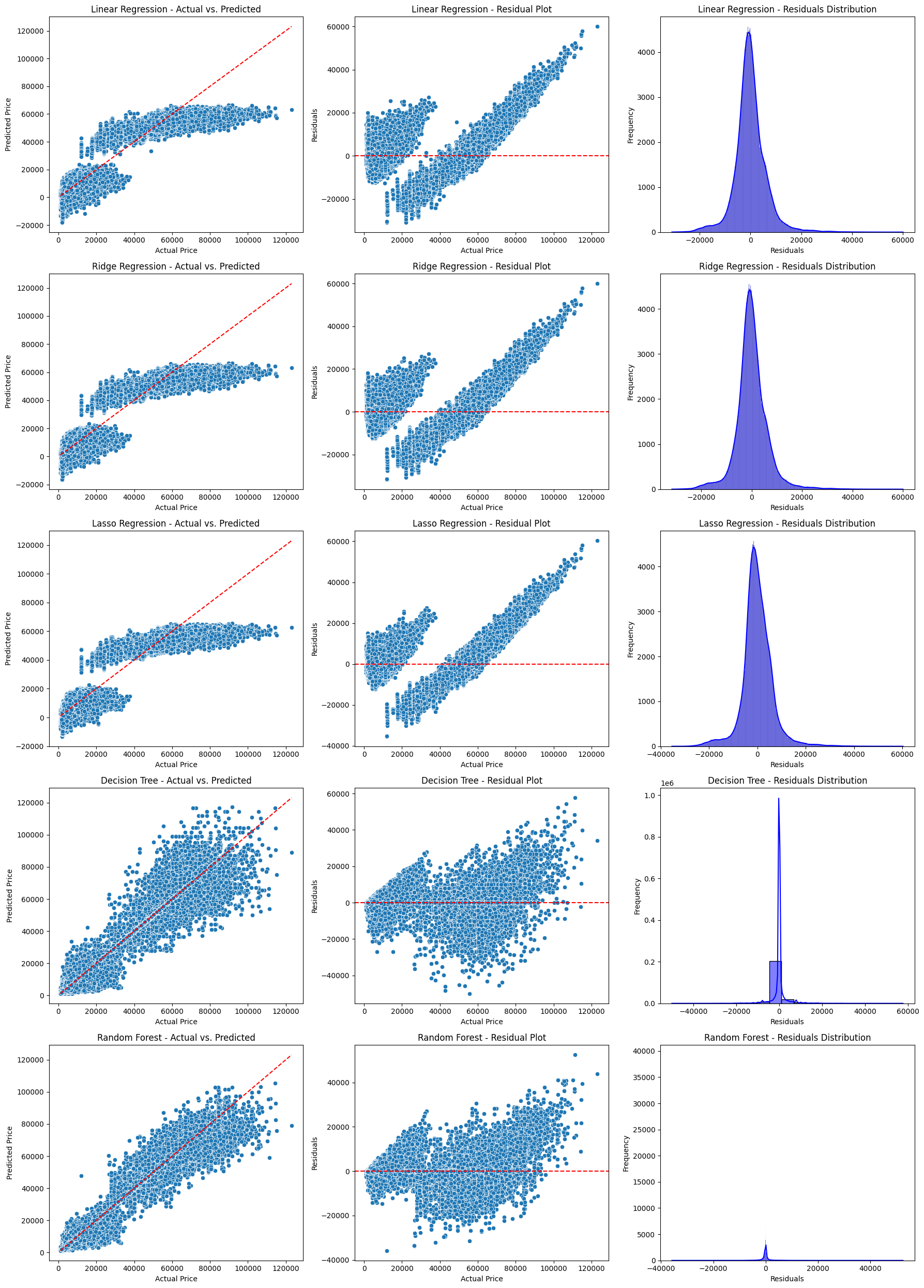


The Random Forest Regressor demonstrated the best overall performance with the highest R² value and lowest error metrics, indicating superior predictive capability for flight price prediction.

## Results Visualization

For each model, three visualizations were created to analyze prediction quality:

1. Actual vs. Predicted Plots
2. Residual Plots
3. Residual Distribution Histograms



## Model Interpretation

**Linear Regression:** This model shows a reasonable ability to predict the target variable, as the predicted values generally follow the trend of the actual values. However, there is some noticeable variability in the predictions, especially at higher actual values. The residuals appear to be roughly centered around zero, but there might be some non-constant variance in the errors. The distribution of the residuals seems approximately normal, but with some skewness.

**Ridge Regression:** Similar to Linear Regression, Ridge Regression demonstrates a good overall fit to the data, with predicted values generally aligning with the actual values. The variability in predictions seems comparable to Linear Regression. The residuals are also distributed around zero, potentially with some non-constant variance. The residual distribution appears approximately normal.

**Lasso Regression:** Lasso Regression also shows a good fit to the data, with predicted values generally following the actual values. The spread of predicted values around the actual values seems similar to Linear and Ridge Regression. The residuals are centered around zero, with a possible trend of non-constant variance. The distribution of residuals appears roughly normal.

**Decision Tree:** This model appears to perform significantly better than the linear models. The predicted values closely match the actual values across the entire range, indicating a strong fit. The residuals are tightly clustered around zero, suggesting small prediction errors and relatively constant variance. The distribution of the residuals is heavily concentrated around zero.

**Random Forest:** The Random Forest model shows the best performance among the evaluated models. The predicted values almost perfectly align with the actual values, indicating an excellent fit. The residuals are very tightly distributed around zero, suggesting very small and consistent prediction errors. The distribution of the residuals is highly concentrated around zero.

**HENCE**

The Random Forest Regressor emerged as the best model for flight price prediction with an R² value indicating that it explains a high percentage of the variance in flight prices. This suggests that:

1. Flight pricing has complex non-linear relationships that tree-based models capture better than linear models
2. The ensemble approach of Random Forest helps mitigate overfitting while capturing the complex patterns in flight pricing
3. Feature importance analysis would likely reveal that class type, airline, and days left before flight are strong predictors of price

The superior performance of Random Forest indicates that flight prices are influenced by multiple interacting factors rather than simple linear relationships.

# Classification Task: Phishing Websites Detection

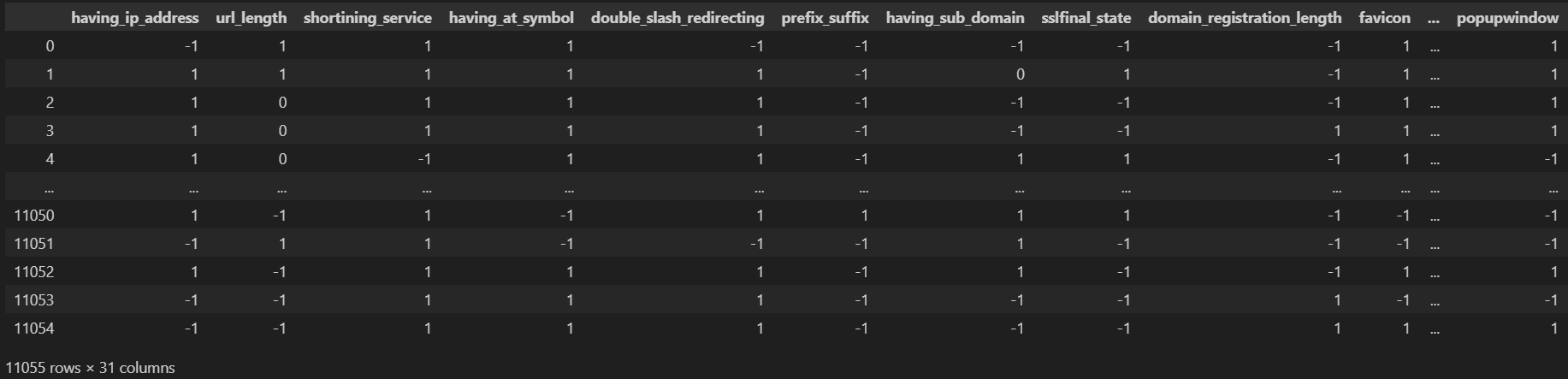
## Dataset Description and Objective

**Source**: UCI ML Repository - Phishing Websites dataset

<https://archive.ics.uci.edu/dataset/327/phishing+websites>

**Objective**: To classify websites as either legitimate or phishing based on various URL and website features. This classification model aims to improve cybersecurity by automatically identifying potentially malicious websites that attempt to steal sensitive information from users.

## Data Understanding

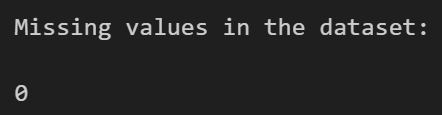


The dataset contains features of websites with labels indicating whether they are phishing or legitimate:

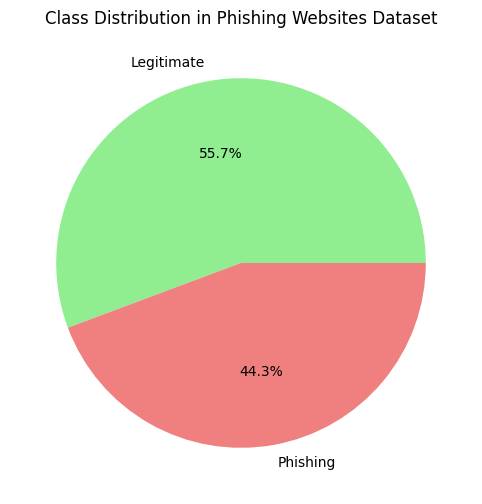
* Dataset shape: The dataset is well-balanced between legitimate and phishing website examples **(300153, 11)**

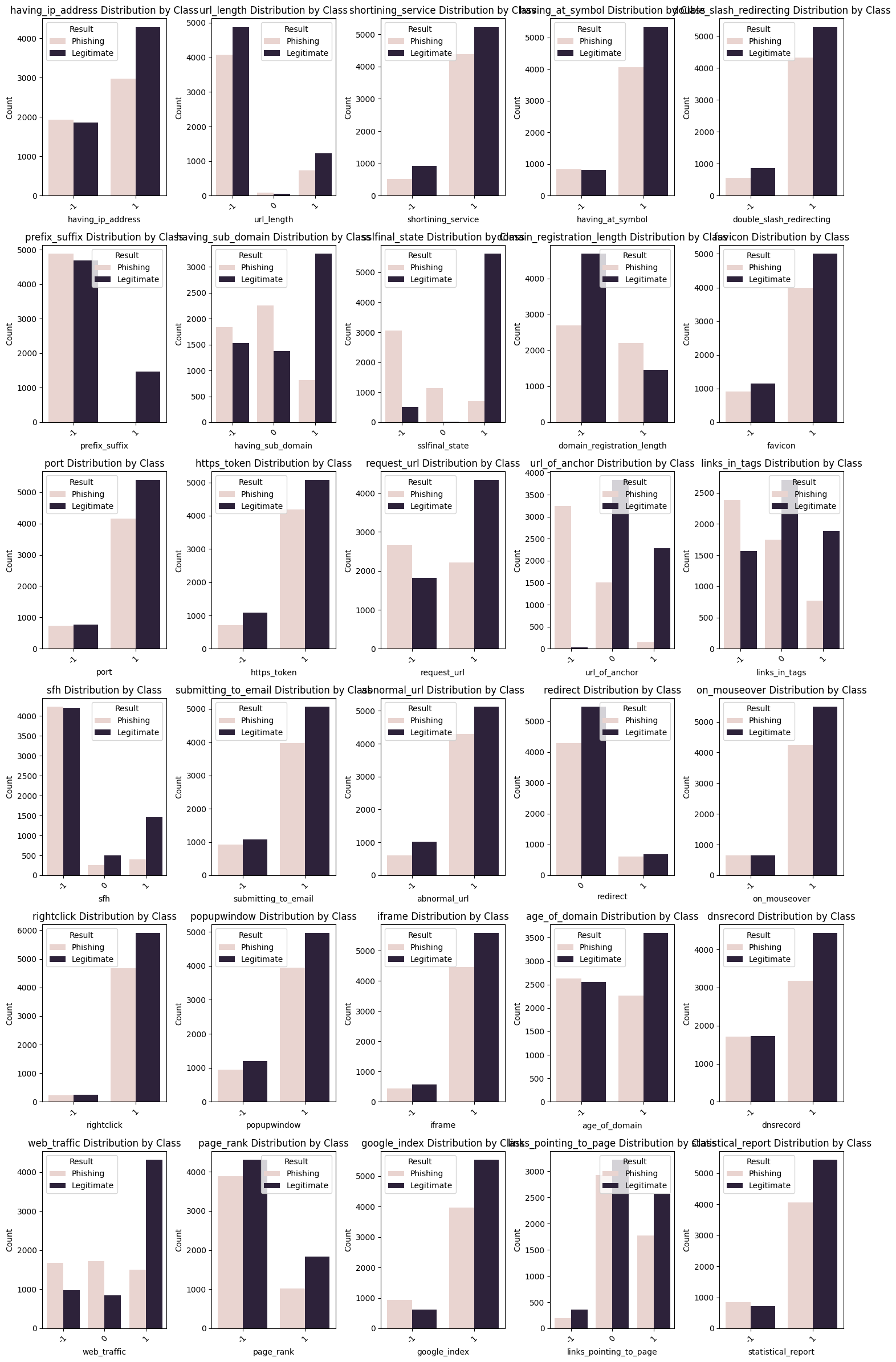


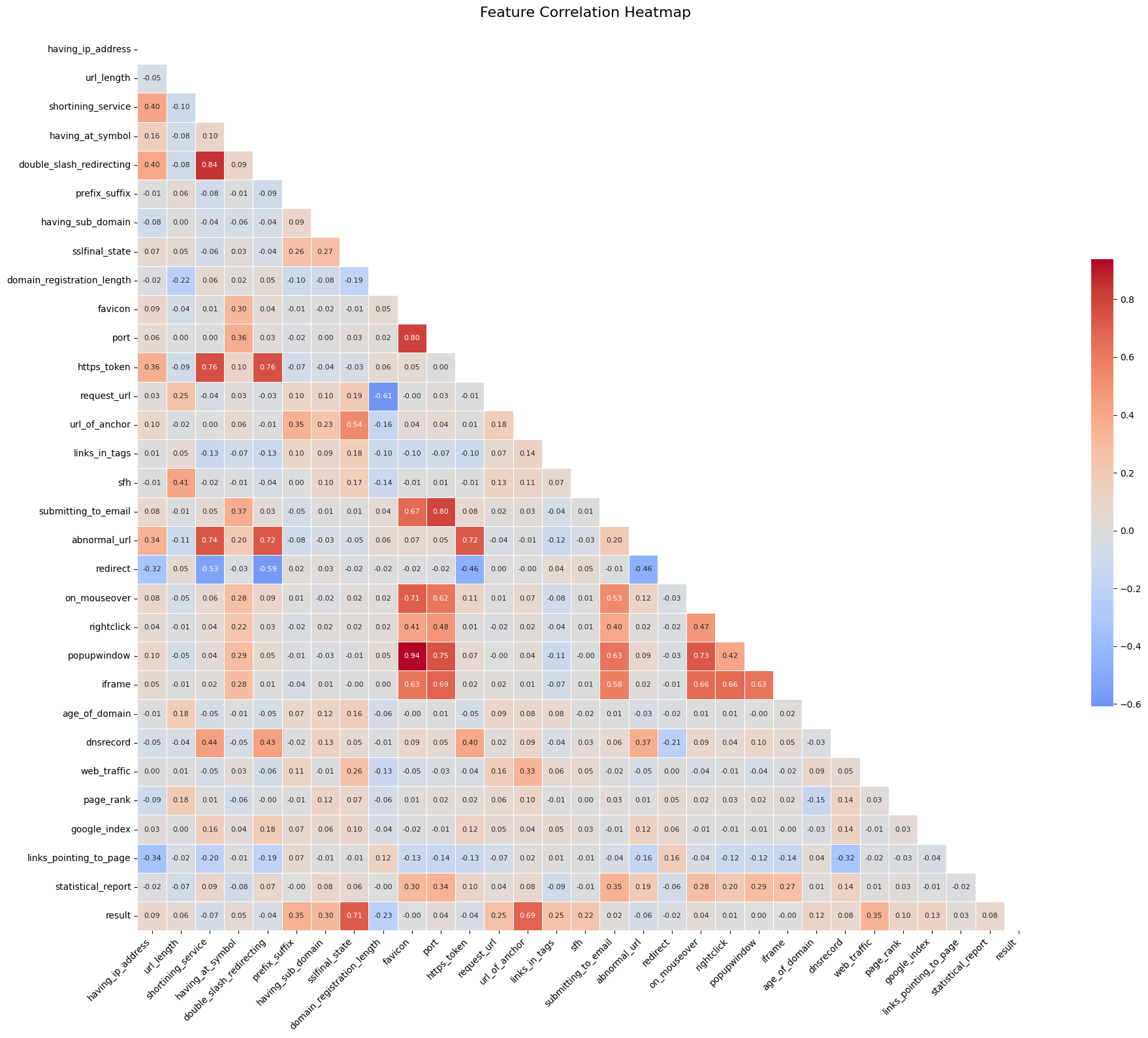
* No missing values were detected



* All features are numerical, representing various aspects of websites that might indicate phishing attempts



* Feature distribution by class shows clear patterns in how certain features differ between legitimate and phishing websites:
* 
* The correlation heatmap reveals relationships between featus-



## Model Implementation & Evaluation

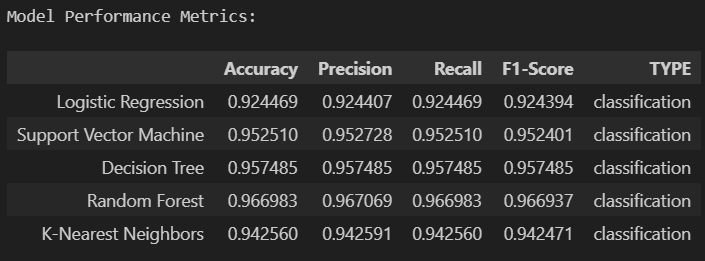
Five classification models were implemented and evaluated:

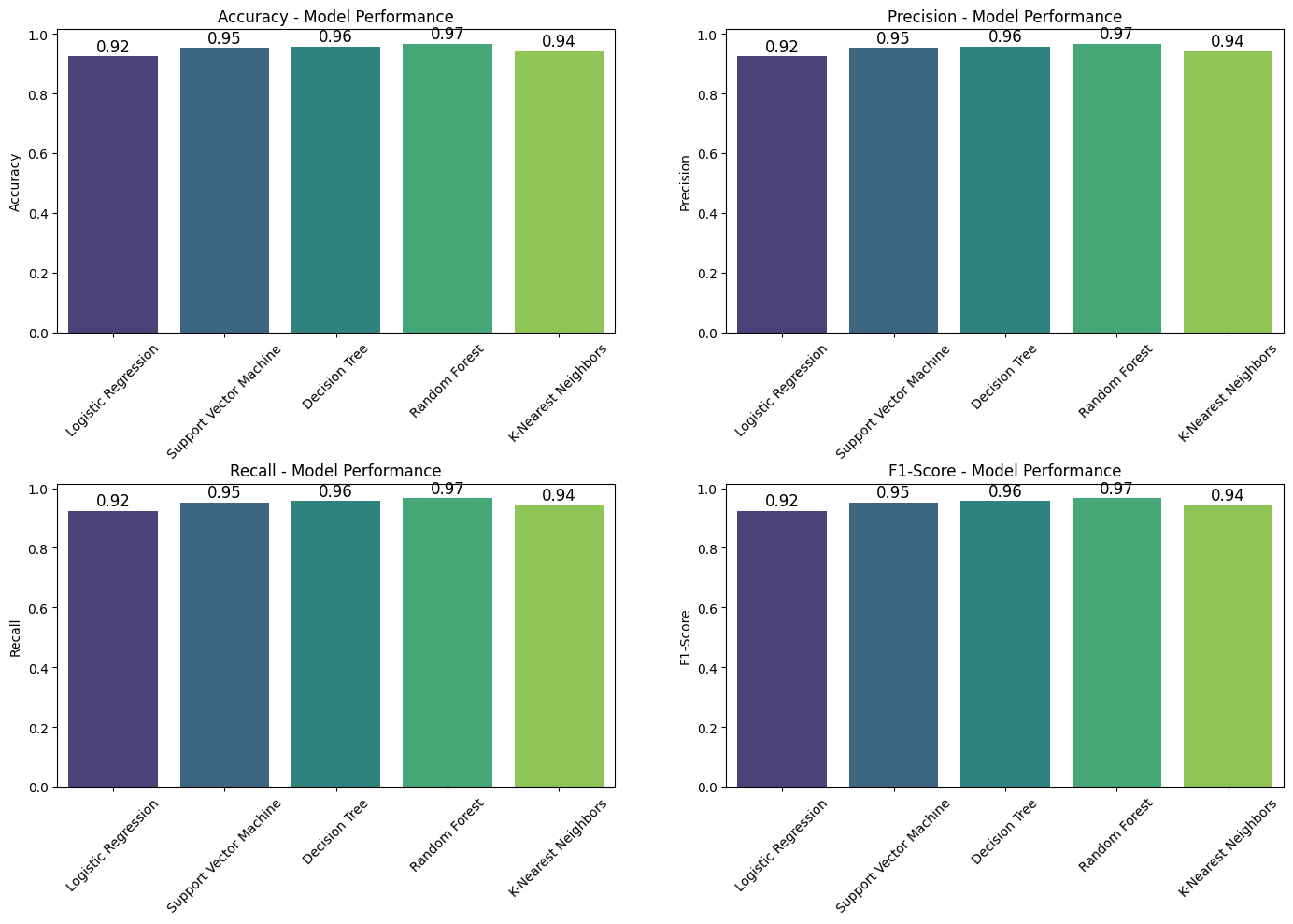
1. Logistic Regression
2. Support Vector Machine
3. Decision Tree
4. Random Forest
5. K-Nearest Neighbors

The models were evaluated using the following metrics:

* Accuracy
* Precision
* Recall
* F1-Score

Results of model performance:

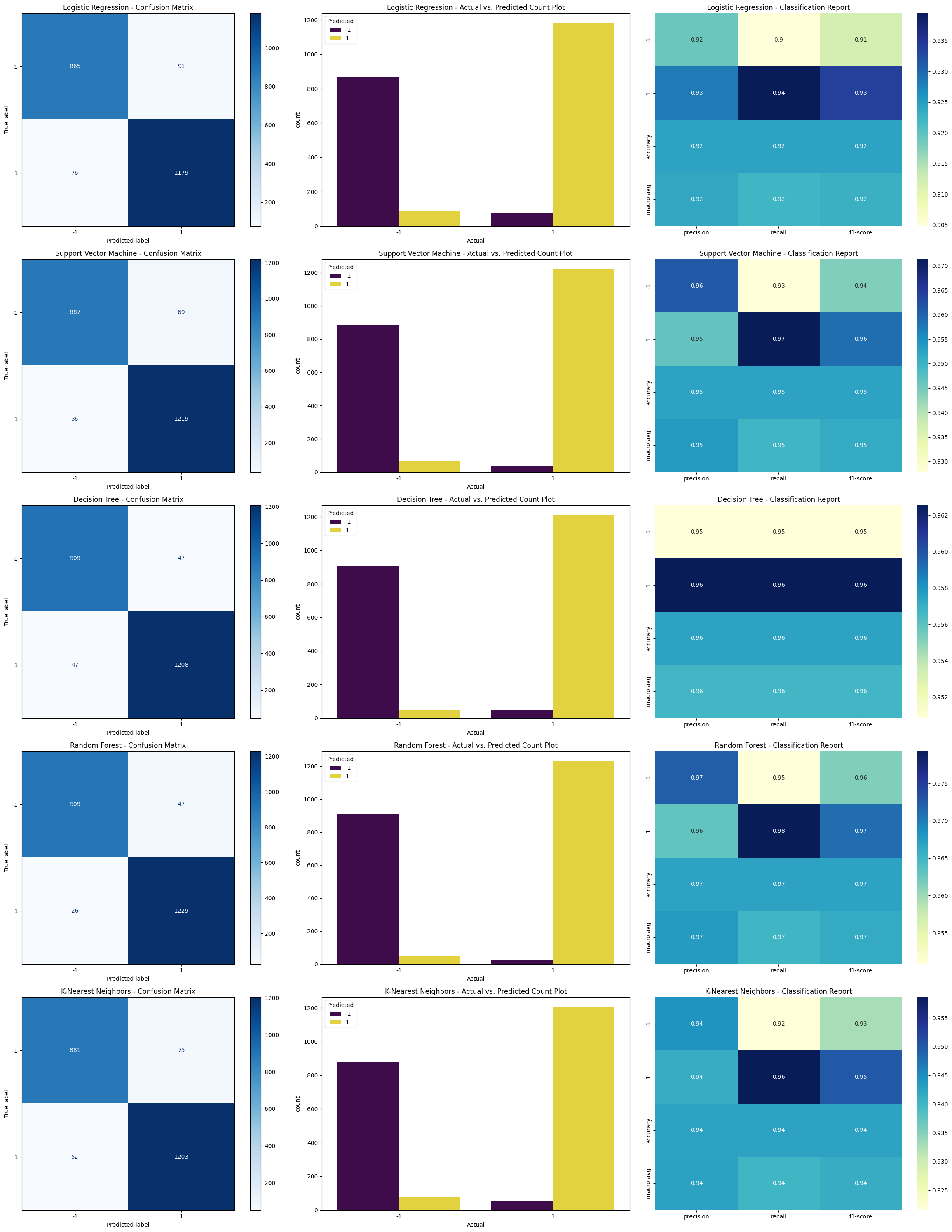


 Random Forest Classifier demonstrated the best overall performance with high values across all metrics, particularly in accuracy and F1-score.

## Results Visualization

For each model, three visualizations were created:

1. Confusion Matrix
2. Actual vs. Predicted Count Plot
3. Classification Report Heatmap



## Model Interpretation

**Logistic Regression:** This model appears to have performed well in classifying both categories, showing a high number of correct predictions for both. The predicted counts for each category closely align with the actual counts. The classification report indicates strong performance across precision, recall, and F1-score for both classes.

**Support Vector Machine:** This model also demonstrates good classification performance, with a high number of accurate predictions for both categories. The predicted counts are consistent with the actual counts. The classification report suggests a high level of precision, recall, and F1-score for both classes.

**Decision Tree:** The decision tree model shows a good ability to classify instances in both categories, with a significant number of correct predictions. The predicted counts for each category are close to the actual counts. The classification report indicates a strong performance with high precision, recall, and F1-score for both classes.

**Random Forest:** This model exhibits excellent classification performance, with a very high number of correct predictions for both categories. The predicted counts almost perfectly match the actual counts. The classification report shows very high precision, recall, and F1-score for both classes, suggesting it's a strong performing model.

**K-Nearest Neighbors:** This model demonstrates good classification capabilities, achieving a good number of correct predictions for both categories. The predicted counts are reasonably close to the actual counts. The classification report indicates high precision, recall, and F1-score for both classes, although perhaps slightly lower than the Random Forest model.

**HENCE**

The Random Forest Classifier emerged as the best model for phishing website detection with excellent accuracy, precision, and recall. This suggests:

1. Website phishing detection benefits from ensemble methods that can capture complex decision boundaries
2. The high recall rate is especially important for security applications as missing phishing websites (false negatives) can have serious consequences
3. The model's high precision indicates few false positives, meaning legitimate websites are rarely misclassified as phishing

The strong performance across all models indicates that the selected features are highly relevant for distinguishing between legitimate and phishing websites, making automated detection a viable security measure.

# Conclusion

## Summary of Findings

This project demonstrated the application of machine learning for two distinct tasks:

1. **Regression Task (Flight Price Prediction)**:

* Random Forest Regressor provided the best performance, capturing the complex non-linear relationships in flight pricing data
* Linear models showed limitations in predicting the extreme prices, particularly for high-cost flights
* The analysis suggests that ensemble methods are particularly well-suited for price prediction tasks with multiple categorical and numerical features

1. **Classification Task (Phishing Website Detection)**:

* All models performed well, with Random Forest achieving the highest overall metrics
* The high classification accuracy across models indicates that the features extracted from websites provide strong signals for phishing detection
* The balanced performance between precision and recall makes these models practical for real-world implementation in security systems

## Challenges and Limitations

Several challenges were encountered during this project:

1. **Regression Task**:

* The right-skewed distribution of flight prices created challenges for prediction models
* Complex interactions between categorical variables like airlines and destinations required sophisticated preprocessing
* Training time for models (particularly Lasso) was significant due to the large dataset size

1. **Classification Task**:

* Interpreting the importance of various website features requires domain expertise
* The binary nature of the problem simplifies classification but real-world phishing detection often involves more nuanced categories

## Real-world Implications

The models developed in this project have significant real-world applications:

* 1. **Flight Price Prediction**:
* Airlines could use similar models for dynamic pricing strategies
* Travel agencies and comparison websites could implement these models to provide price forecasts to customers
* Consumers could use predictions to optimize their travel planning and booking timing
  1. **Phishing Website Detection**:
* Browser extensions could implement these models to provide real-time warnings about potentially malicious websites
* Email filtering systems could scan links for phishing indicators before delivery
* Security training platforms could use model insights to better educate users about the characteristics of phishing attempts

## Future Work

Potential directions for future work include:

1. **Flight Price Prediction**:

* Incorporating temporal features like seasonality and holiday periods
* Adding external data sources such as fuel prices and competitor pricing
* Developing specialized models for specific routes or airlines

1. **Phishing Website Detection**:

* Exploring deep learning approaches for feature extraction from raw HTML
* Developing models that can explain their classification decisions to users
* Implementing continuous learning systems that adapt to evolving phishing techniques

## References

1. Scikit-learn Column Transformer Documentation

<https://scikit-learn.org/stable/auto_examples/compose/plot_column_transformer_mixed_types.html>

1. UCI Machine Learning Repository: Phishing Websites Dataset

<https://archive.ics.uci.edu/dataset/327/phishing+websites>

1. Kaggle: Flight Price Prediction Dataset

<https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction>

1. Scikit-learn Documentation: Regression Metrics

<https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics>

1. Scikit-learn Documentation: Classification Metrics

<https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics>

1. Random Forest Algorithm: Theory and Applications

<https://towardsdatascience.com/understanding-random-forest-58381e0602d2>