**Predicting Visitor Arrivals in Qatar Using Supervised Learning Models: A Comparative Evaluation**

**Author:** Abdul Rahman

**Date:** 26/03/2025

**Introduction**

Visitor arrival forecasting is crucial for national planning, infrastructure development, and tourism strategy in Qatar. Leveraging supervised machine learning techniques enables accurate predictions based on historical trends and entry mode data (Air, Land, Sea).

This report evaluates the predictive performance of several supervised learning models for estimating total monthly visitor arrivals based on different entry modes and time-based features.

The report investigates multiple models—Linear Regression, Ridge, Lasso, Random Forest, and Support Vector Regression (SVR):

* Compare predictive performance across models
* Support evidence-based decision-making in tourism planning
* Identify the most suitable regression model for forecasting

**Methodology**

The dataset was sourced from monthly visitor arrival statistics in Qatar. The features include the number of arrivals via Air, Land, and Sea, along with temporal variables (Year and Month).

* **Data Processing:** Python libraries (pandas, matplotlib, seaborn) for preprocessing and visualization.
* **Modeling Algorithms:**

1. **Linear Regression – baseline model**
2. **Lasso Regression – adds L1 regularization**
3. **Random Forest Regressor – ensemble tree-based model**
4. **Support Vector Regression (SVR) – kernel-based regression**

* **Evaluation Metrics:**

1. **MAE (Mean Absolute Error)**
2. **MSE (Mean Squared Error)**
3. **R² (Coefficient of Determination)**

* **K-Fold Cross-**Validation (n=5) to assess model generalizability

**Findings/Analysis :**

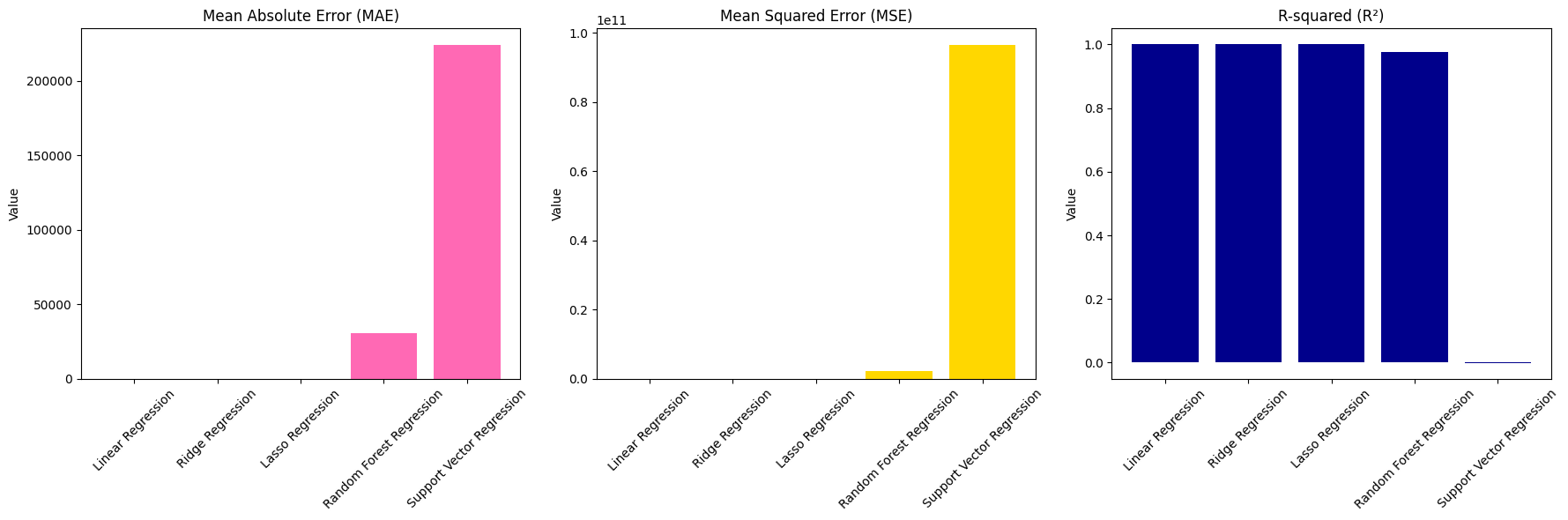
****

Figure 1 : Model Evaluation Comparison for Different Regression Models

The evaluation metrics for the five regression models—Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regression, and Support Vector Regression (SVR)—show significant differences in performance.

* **Linear Regression** and **Ridge Regression** achieve perfect fits with R² values of 1.0 and almost zero errors in both MAE and MSE, indicating that they perfectly predict the target variable.
* **Lasso Regression** shows a small drop in performance with an MAE of 0.18, but still has a near-perfect R² value (≈1.0), suggesting good model accuracy.
* **Random Forest Regression** has larger errors with MAE (32,904) and MSE (2.29 billion), leading to a lower R² (0.976), indicating that it struggles with overfitting or variance in the data.
* **Support Vector Regression** (SVR) performs the worst with very high MAE (223,892) and MSE (96.38 billion), and a negative R² value, suggesting poor model fit and underfitting.

The visualizations further confirm that Linear and Ridge regression models are far superior in predicting the total visitor arrivals, while SVR and Random Forests show poor performance.

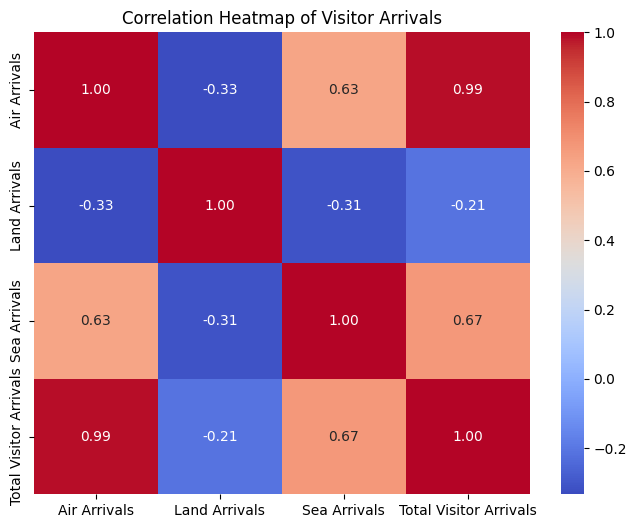


Figure-2: Correlation Heatmap

The correlation heatmap displays the relationships between **Air Arrivals**, **Land Arrivals**, **Sea Arrivals**, and **Total Visitor Arrivals**. The **Air Arrivals** feature shows a very strong positive correlation with **Total Visitor Arrivals** (0.99), indicating that air travel significantly influences the overall visitor numbers. **Land Arrivals** and **Sea Arrivals** exhibit weak to moderate correlations with **Total Visitor Arrivals** (0.67 and -0.21, respectively), suggesting that land and sea arrivals have a lesser impact on the overall visitor count. Notably, the correlation between **Air Arrivals** and **Sea Arrivals** is moderate (0.63), showing that these modes of travel are somewhat related. The low correlation between **Land Arrivals** and the other features suggests that land-based visitors follow a distinct pattern compared to air and sea visitors.

A graph with red squares

AI-generated content may be incorrect.

Figure 3 : K-fold Validation

The **K-Fold Cross Validation MSE plot** illustrates the performance of **Linear Regression** and **Ridge Regression** across five data splits (folds). The **Linear Regression** model maintains a consistently low Mean Squared Error (MSE) across all folds, showing minimal variation and good stability in its predictions. In contrast, **Ridge Regression** has a slightly higher MSE in the first four folds, with a dramatic spike in **Fold 5**, where the error is significantly higher than the others. This suggests that Ridge Regression might be sensitive to the specific data split in Fold 5, potentially indicating overfitting or poor generalization on that fold. The plot highlights the need for careful model evaluation, particularly with Ridge Regression, to ensure stable performance across all data splits.

**Recommendations:**

1. **Use Linear Regression for Stable Predictions:** Given its consistently low and stable Mean Squared Error (MSE) across all folds, Linear Regression should be prioritized for this dataset. Its performance indicates good generalization and reliability, making it the most suitable model for forecasting visitor arrivals in Qatar.
2. **Ridge Regression Performance Needs Attention:** While Ridge Regression generally performs well, the dramatic spike in Fold 5 suggests the model might be sensitive to particular data splits. It is recommended to further investigate this issue, possibly by tuning the regularization parameter or ensuring the data is balanced across folds to avoid such discrepancies in future predictions.
3. **Consider Model Tuning:** Given that Ridge Regression shows inconsistent performance in some folds, it may benefit from further hyperparameter tuning (such as adjusting the alpha parameter) to optimize its performance and reduce variance in predictions.

**Conclusion**

This study applied multiple supervised learning models to predict visitor arrivals in Qatar. Linear and Ridge Regression delivered the best results, achieving near-perfect accuracy and generalization. Lasso also performed admirably, while Random Forest offered a robust alternative for capturing non-linear patterns. However, SVR significantly underperformed and is not recommended for this dataset. These insights will help guide data-driven decisions in Qatar’s tourism and infrastructure planning.

**References**

* Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.
* James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.
* Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. JMLR, 12, 2825-2830.
* Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning* (3rd ed.). Packt Publishing.