**Predicting Airport Efficiency and Delays: A Linear Regression Approach applied to the case of EWR International Airport.**

Sreya Chowdary and Uday Chaitanya Padala

Department of Data Science, University of Maryland Baltimore County DATA 606: Capstone in Data Science

Dr. Tony Diana, May 04, 2024

# **Abstract**

This study compares the performance of various machine learning models to predict taxi-in delays at Newark Liberty International Airport (EWR), and data is sourced from the FAA's ASPM database. The methodology incorporated data pre-processing techniques like cleaning and normalization alongside feature engineering to refine the dataset for advanced analysis and compared the performance of linear regression, lasso, ridge, random forest, and XGBoost models. The objective was to identify the model that best describes and predicts operational inefficiencies contributing to delays. Lasso and ridge are better in model interpretability, and XGBoost appears to be the best model with the lowest root mean squared error and the highest R-squared. It offers the highest predictive accuracy and explains the most variance in taxi-in delays.

(KEYWORDS: Machine learning, Predictive modeling, Regularization, Ensemble methods)

1. **Introduction**

In 2019, Newark Liberty International Airport (EWR) served over 46 million passengers. However, the summer of 2023 saw significant delays at EWR, highlighting the need for a comprehensive assessment of airport operations, particularly taxi times, which is a critical indicator of tarmac congestion (Park, D. K., & Kim, J. K., 2018). Recent studies, such as those by Anupkumar (2023), have also emphasized the impact of increasing passenger numbers on airport operational metrics, calling for an urgent evaluation of operational efficiency.

This study aims to investigate the factors contributing to operational delays at EWR, with a primary focus on taxi times, to enhance airport efficiency and passenger experience. Extended taxi times at EWR often indicate inefficiencies in aircraft sequencing and gate management, reflecting concerns regarding airport capacity (Lee, M. K., & Park, D. K., 2021). By addressing these inefficiencies, this research aims to contribute to a better understanding and improvement of airport operations.

Understanding the paper of Diana T in 2018, where he emphasized the importance of taxi-out-delay metric and Predicting taxi-out times is of great interest to airport, airline, and regulatory analysts to ensure on-time departure performance. So, in his paper he tried to see if ensemble models are predicting better than predictive models. I propose an intriguing test to explore this further, considering average-taxi-in-delay as the predictor variable. This variable is a key factor for inefficiencies in aircraft maintenance, making the test even more compelling.

To achieve this goal, the study employs a combination of statistical and machine learning models, including Linear Regression, Lasso and Ridge Regression, Random Forest, and XGBoost, to predict and assess taxi times. Each model is selected for its unique ability to handle the complexities of regression analysis in operational scenarios, as discussed in the comprehensive machine learning literature by Géron (2017). Through rigorous evaluations utilizing real-time data and predictive analytics, the project aims to enhance strategic planning and operational decision-making at EWR.

The study will delve into key performance indicators, such as taxi-in delays, gate management, and overall on-time performance, utilizing hourly operational metrics from the FAA’s ASPM database for the years 2018 and 2023. By analyzing historical taxi-in delay data, the research aims to identify key predictors of delays and develop recommendations for airport management to streamline ground operations and improve overall airport performance (Zhang, L., & Wang, Y., 2010).

1. **LITERATURE REVIEW**

Recent advancements in airport operations research have underscored the critical importance of optimizing taxi-in and taxi-out times to enhance overall airport efficiency. Taxi-in time, the duration an aircraft spends taxiing from the runway to its designated gate, is influenced by various factors such as airport layout, air traffic congestion, and gate availability. Efficient management of both taxi-in and taxi-out times is essential for minimizing operational delays and improving passenger satisfaction.

According to a study conducted by Diana T in 2018, taxi times are crucial metrics for analyzing flight schedules, fuel consumption, and passenger satisfaction. It is also important for airport operations and planning to identify choke points between gate-out and take-off to recommend and implement infrastructure improvements for optimal surface movement flows. Airport operators can estimate departure clearance time compliance. The study evaluated whether ensemble learning models are more likely to improve forecasts of taxi-out times than traditional linear models.

Studies conducted by Li et al. (2020), Park & Kim (2023), and others have elucidated the intricate relationship between airport layout complexity and traffic congestion in both taxi-in and taxi-out processes. They highlight how convoluted runway and taxiway configurations necessitate shared paths for arriving and departing aircraft, leading to increased congestion and prolonged taxi times. These studies advocate for strategic gate assignments, improved ground-handling efficiencies, and optimizing departure flows to mitigate such challenges effectively.

Traditional linear models have shown limitations in capturing the nuanced dynamics of airport operations in both taxi-in and taxi-out scenarios. Prompted by this, researchers have increasingly adopted advanced predictive models, including machine learning techniques such as Random Forest and XGBoost. Choi and Kim (2021), Shaik and Teja (2019), and others emphasize the efficacy of these models in predicting and managing airport operational efficiencies, citing their capacity to accommodate nonlinear interactions and identify significant predictors of taxi-in and taxi-out delays.

**Table 1: Comparision of different studies on taxi times**

|  |  |  |  |
| --- | --- | --- | --- |
| **Author(s)** | **Research Focus** | **Contributions** | **Implications for the Study** |
| Yin et al. (2024) | Advanced machine learning frameworks for predicting taxi times | Supported the use of complex AI techniques, such as convolutional neural networks, for enhancing accuracy. | Suggests integrating advanced analytical models to improve the accuracy of predicting taxi times at EWR. |
| Ravizza et al. (2014) | Comparison of regression algorithms for modeling nonlinear data | Recommended TSK fuzzy rule-based systems for their efficacy in handling diverse airport operations. | Advises on selecting appropriate modeling techniques to address the complexities of airport operations. |
| Federal Aviation Administration (FAA) | Operational metrics and definitions relevant to airport operations | Provided detailed definitions and datasets for operational metrics such as taxi-in and taxi-out times. | Ensures accurate usage of operational metrics in the study, facilitating precise analysis and recommendations. |

The impact of delays and inefficiencies in both taxi-in and taxi-out times extends beyond operational challenges, significantly affecting airport economic performance and passenger satisfaction. Efthymiou et al. (2018), Anupkumar (2023), and others quantify these impacts, revealing substantial annual losses due to inefficiencies. They underscore the importance of efficient airport operations in maintaining economic viability and customer trust.

Scholarly inquiry into predictive models and methodologies for anticipating airport surface taxi times has further enriched the field. Yin et al. (2024) advocate for advanced analytical frameworks incorporating machine learning and simulation models to enhance predictive accuracy. Ravizza et al. (2014) compare multiple regression algorithms and recommend TSK fuzzy rule-based systems for their superior accuracy in modeling nonlinear data patterns, which is particularly beneficial for airports with diverse operating characteristics.

In conclusion, this comprehensive literature review provides insights into recent advancements in optimizing both taxi-in and taxi-out times in airport operations. It shows the significance of efficient management strategies and need for advanced predictive modeling techniques to enhance passenger experience and overall airport efficiency. However, challenges such as interpretability, data volume requirements, and model complexity persist, highlighting the need for continued research and validation to improve predictive accuracy within the dynamic airport environment.

## **3. Methodology**

### **3.1 Data Understanding**

The data, derived from the Aviation Systems Performance Metrics (ASPM) data warehouse, equips us with the necessary insights on operations and delays. This includes key metrics such as taxi times, delays, on-time performance, efficiency, and challenges faced across different times of the day and year. By organizing the data by hour, we can conduct a proactive analysis of peak vs. off-peak performance, identify specific hours prone to greater delays, and adjust operational strategies accordingly, ensuring a smoother future for the aviation industry.

Data provides a longitudinal view, with where we can compare the metrics between two different years before and after covid-19, 2018 and 2023. This comparison is beneficial for identifying the impact of operational changes, efficiency initiatives, and external factors over time. Perform trend analyses, evaluate the effectiveness of implemented policies or changes, and forecast future trends based on historical data.

The datasets, airport\_efficiency, and airport\_analysis, were cleaned and merged based on the 'Local\_Hour' criterion. Subsequently, the combined dataset is checked for null values and duplicates to ensure the accuracy and integrity of the data. The final dataset consisted of 576 rows and 36 columns. These steps were crucial for maintaining the quality and reliability of the dataset, which underpins the robustness of the subsequent analysis presented in this study.

The final dataset has several variables, that potentially influences delay times Calendar\_year and local \_hour offers temporal context, Departures and arrivals for efficiency computation reflect operational volume, On-time performance percentages with respect to operations and Various average delay metrics (gate departure, taxi-out, airport departure, airborne, taxi-in, block, and gate arrival) provide detailed delay profiles.

**3.2 Response variable**

Average Taxi-in Delay is the response or dependent variable, it is a critical metric used to assess the efficiency of the airport's ground operations. It measures the time aircraft spend taxiing to the gate after landing until they complete stop at the designated gate. This duration can be influenced by several factors such as airport layout, traffic congestion, gate availability and operational policies (ACI).

This variable is essential for identifying inefficiencies in airport ground operations which can significantly impact passenger satisfaction and operational costs for airlines. By analyzing average taxi-in delay, airport management can devise strategies to streamline operations, improve punctuality, and enhance the overall travel experience.

**3.3 key performance indicators of operations research**

When considering average taxi-in delay as the dependent variable in airport operations research, it is crucial to understand and consider the factors that can influence changes in taxi-in times.

***Scheduled Arrivals*** indicates the volume of air traffic an airport handles within a specific hour, it is one of the factor that can directly impact taxi-in delays due to congestion on taxiways or limited gate availability.

***Average Gate Arrival Delay***is the delay experienced by arriving flights from their scheduled gate arrival time, offering insights into potential inefficiencies or operational challenges that also affect taxi-in times.

***Average Taxi Out Time and Delay*** is the indirect feature that can impact, particularly in single-runway or limited-taxiway environments where arriving and departing flights share routes.

***% On-Time Gate Arrivals***indicates the efficiency of flights arriving on schedule. A lower percentage could signal bottlenecks in ground operations that also affect taxi-in durations.

***Facility Reported Arrivals*** reported by the facility can differ from scheduled data, influencing ground operations management and resulting taxi-in delays.

***Departure Efficiency and Arrival Efficiency*** measures how well departures and arrivals are managed with their scheduled times, indirectly impacting the taxi-in times as delays in one part of the operation can cascade to others.

**3.4 Exploratory data analysis**

This dataset includes a wide range of features that provide insights into airport efficiency, on-time gate departures and arrivals, and capacity utilization. To ensure the analysis is consistent and dependable, the data underwent preprocessing, which involved cleaning missing values, normalizing data formats, and deriving new variables to enhance the models' predictive capabilities.

The dataset primarily comprises numerical features, which require several preprocessing steps to ensure optimal conditions for analysis. Key preprocessing activities included are

***Conversion of Numerical Data*** columns containing string representations of numbers were converted to appropriate numeric data types. This transformation is crucial for enabling accurate mathematical operations and statistical analysis.

***Handling of Duplicate Values*** within the dataset was identified and removed to prevent skewing of results and to ensure the integrity of the analysis.

***Imputation of Missing Values***helps maintain the dataset’s robustness, allowing for more reliable and consistent analytical outcomes.

***Creating Time-Based Features*** like 'Month' and 'Dayofweek' are derived from the date data. These features help analyze trends and patterns across different timescales, which can be crucial for understanding seasonal variations and weekly cycles in airport operations. Incorporating time-based features for analyzing airport operations provides multiple advantages. It allows for a better understanding of temporal patterns, enhances the predictive performance of models, and supports more effective operational and strategic decision-making. These features are essential for managing the complex dynamics of airport traffic and operational efficiency.

The data underwent preprocessing stages to prepare it for analysis and modeling.

From Scatter plots, we can say Like in 2023, there is another increase in delays during the evening hours from 4 PM to 8 PM, suggesting that evening rush hours also lead to increased delays. However, the peak delays in 2018 appeared earlier, around 6 PM to 7 PM, compared to 8 PM in 2023.

**Distribution of key variables**

The variables in our dataset exhibit diverse distribution patterns, which are essential for determining the appropriate preprocessing and analytical methods.

**Table 2: Understanding the distribution of features**

|  |  |
| --- | --- |
| **Distribution Type** | **Variables** |
| **Normal** | On-time airport departures, % on-time gate departures, % capacity utilized |
| **Uniform** | Calendar year, Date |
| **Right-Skewed (R)** | Facility Reported Departures, Facility Reported Arrivals, Total Facility Reported, Departure Demand Units, Arrival Demand Units, Total Demand Units, Average Taxi In Delay, Average Taxi Out Time, Average Taxi Out Delay, Average Gate Arrival Delay, Average Airport Departure Delay, Average Airborne Delay, Capacity AAR, Departure Efficiency, Arrival Efficiency |
| **Left-Skewed (L)** | % On-Time Gate Arrivals, Average Gate Departure Delay, Average Block Delay |
| **Not Categorized** | Departure for efficient computation, Total facility reported, Average taxi delay, Departure demand unit, Average taxi time |

Variables such as 'on-time airport departures' and '% on-time gate departures' approximate a normal distribution, indicating that the data around the mean are symmetrically distributed without significant skew.

Right-Skewed variables are 'facility reported departures', 'average taxi in delay', 'average taxi-out time' and 'average airport departure delay are crucial for analyzing airport capacity, studying ground handling efficiency, and the impact of airfield congestion.

By categorizing our variables according to their distribution, research for deeper insights and more robust statistical analysis, supporting stronger conclusions and potentially more impactful results.

**Feature scaling**

The dataset comprised numerous numerical columns such as 'Average\_Taxi\_Out\_Time', 'Departure\_Efficiency', and '% On-Time\_Gate\_Arrivals', which represent varied aspects of airport operations. These features had different units and scales; time-related measures were in minutes, while efficiency metrics were percentages.

To prepare the data for predictive modeling, the Min-Max Scaler was applied to these numerical columns after initial data cleaning steps. Data cleaning is done by removing non-numeric characters, converting sting data types to numeric, and missing values are handled by dropping rows to see numerical data was missing. This ensured that the scaling process was accurate and meaningful. The scaled data was then converted to a pandas DataFrame to do further analysis on this data.

Scaling is essential for linear models such as linear regression and Lasso regression since it significantly affects the optimization process and regularization. It guarantees that the model's performance is not affected by the inherent magnitude of features, hence resulting in more precise and dependable predictions. This is particularly crucial in situations where the selection of features and the interpretability of the model are of utmost importance, as is the case in numerous practical applications.

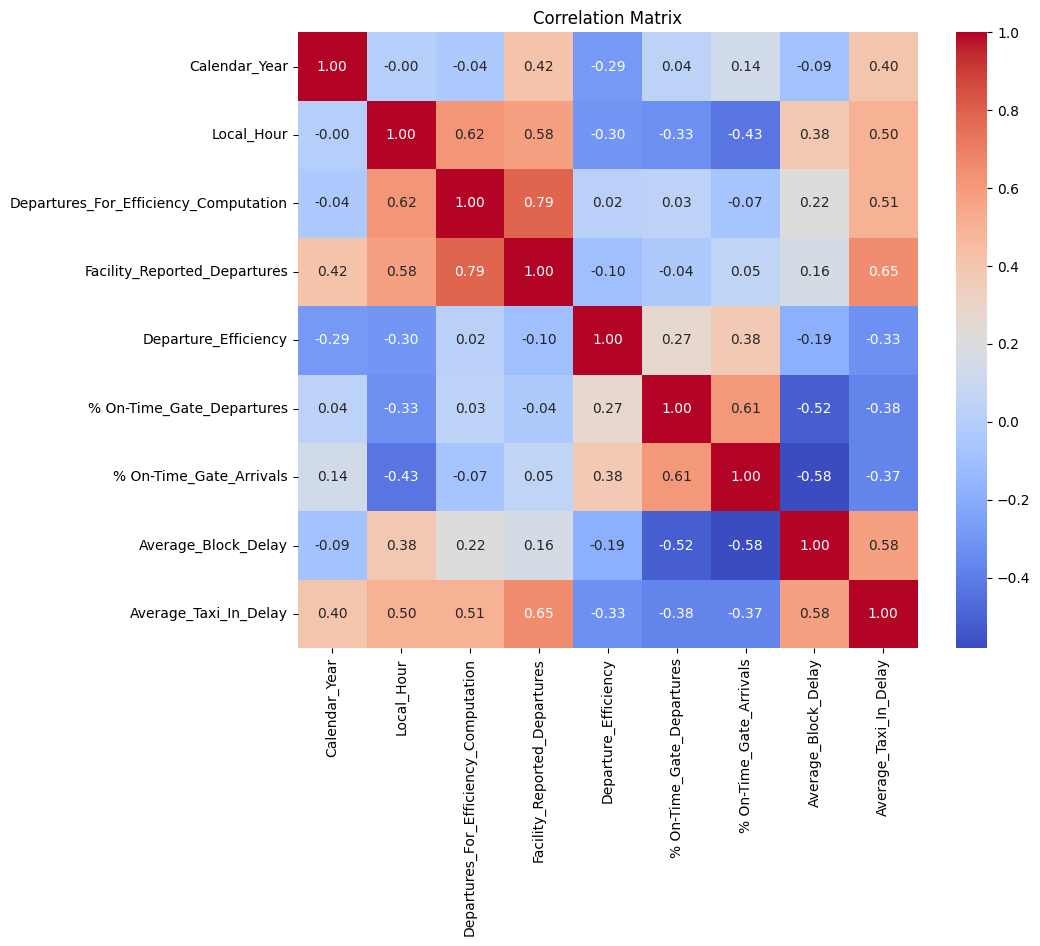
**Multicollinearity**

Since there are many features in the dataset, there is a possibility of multicollinearity. To check for this, I used a correlation matrix, as it is one of the effective variable selection techniques. In this method, the correlation matrix is reviewed, and one of two variables is removed when they exhibit a high correlation—typically greater than 0.8 or another threshold appropriate to the context of the data.

As we are using linear regression, multicollinearity can inflate the variance of the coefficient estimates, leading to unstable models. However, regularization methods, such as Lasso and Ridge, can be beneficial as they can handle multicollinearity to some extent by penalizing the coefficients of correlated features.

Many of the features are closely related to each other, indicating redundancy. For instance, 'Scheduled\_Arrivals' and 'Arrival\_Demand\_Units' are highly correlated, suggesting they may convey similar information. Similarly, 'Capacity\_AAR', 'ADR', 'ADR Plus\_Capacity\_AAR', and 'Average\_Capacity' form another set of closely related features. Consequently, one of two variables is removed when they exhibit high correlation, and the final set of features is derived from this step. After this step, the correlation matrix looked like this

**Fig 1: Correlation matrix for the features after removing multicollinearity**



**STATS model**

STATS models is a library that is specifically created for the purpose of statistical modeling and analysis. The software provides powerful capabilities for estimating diverse statistical models, executing statistical tests, and enabling comprehensive data exploration. This library is highly beneficial for regression analysis, one of its most common uses. It gives detailed statistical data, making it appropriate for traditional statistical analysis and modeling.

### **OLS Regression Analysis Interpretation**

Model Fit and Effectiveness

* + R-squared (0.740) suggests that approximately 74% of the data variation in the value 'Average\_Taxi\_In\_Delay' is explained by the model. Such a high R-squared value indicates a strong model fit to the data.
  + Adjusted R-squared (0.735), After adjusting for the number of predictors, still the value is same indicating it as robust, reaffirming the model’s effectiveness.

Statistical Significance and Coefficients

Each coefficient in the model represents the change in the dependent variable for a one-unit change in the predictor variable, assuming all other variables are held constant.

* Calendar\_Year: Each additional year is associated with an increase of 0.0919 in 'Average\_Taxi\_In\_Delay', significant at the 0.000 level.
* Average\_Block\_Delay, Shows a strong positive relationship with 'Average\_Taxi\_In\_Delay' (coefficient = 0.4741), suggesting that as block delays increase, taxi-in delays also increase significantly.
* % On-Time\_Gate\_Departures and % On-Time\_Gate\_Arrivals: Both variables negatively affect the 'Average\_Taxi\_In\_Delay', indicating that better on-time performance correlates with reduced taxi-in delays.

Model Diagnostics

* F-statistic (160.3): Indicates the model is statistically significant.
* Prob(F-statistic): The very low probability (<0.0001) suggests the model predictors provide a good fit over the null hypothesis.

In the regression model, selecting the features with a p-value less than 0.005 confirms statistical significance. This threshold helps prevent overfitting and enhances the model's predictive accuracy by focusing on predictors that impact the dependent variable. Features such as Calendar\_Year and Average\_Block\_Delay demonstrate robust relationships, indicating their significant influence on Average\_Taxi\_In\_Delay. This approach not only strengthens the theoretical foundations but also enhances practical applications.

Using this method, we filtered the features, and finally, we got seven features that are highly important for predicting average taxi-in delays. The features 'Local\_Hour', 'Calendar\_Year', 'Facility\_Reported\_Departures', 'Departure\_Efficiency', '% On-Time\_Gate\_Departures', '% On-Time\_Gate\_Arrivals', and 'Average\_Block\_Delay' are critical for predicting 'Average\_Taxi\_In\_Delay'. Their significance is underscored by statistical tests and practical implications in operational dynamics. These predictors capture both temporal and operational aspects, to reduce taxi-in delays at airports. Their inclusion enhances model accuracy and, most importantly, provides invaluable insights for airport management.

**4. Modelling**

Predictive models for regression analysis and ensemble models like random forests or gradient boosting are used to forecast taxi-in delays. These models can help airports and airlines in making proactive decisions to manage resources and schedules effectively. Linear regression models are constructed for each delay type as a function of all other available variables. Model selection is guided by statistical significance, coefficient size, and model fit metrics such as R-squared values.

**A. Linear Regression**

The data sample was split into training (80%) and testing (20%) sets. A linear regression model was then trained using the sklearn library. Linear regression is a statistical technique that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The premise is that a change in the independent variables will affect the dependent variable proportionally (Fernandes, 2024).

### **Mathematical Formulation**

In a simple linear regression with one independent variable, the model is typically expressed as:

*𝑌=𝛽0+𝛽1𝑋+𝜖Y*=*β*0 +*β*1 *X*+*ϵ*

* *𝑌Y*: Dependent variable
* *𝛽0*: Intercept of the regression line (the value of *𝑌Y* when *𝑋X* is 0)
* *𝛽1*: Slope of the regression line, which is the coefficient of the independent variable *𝑋X*
* *𝑋*: Independent variable
* *𝜖*: The error term representing the difference between the observed values and the values predicted by the model

For multiple regression involving several independent variables, the equation extends to:

*𝑌=𝛽0+𝛽1𝑋1+𝛽2𝑋2+...+𝛽𝑛𝑋𝑛+𝜖Y*=*β*0 +*β*1 *X*1 +*β*2 *X*2 +...+*βn* *Xn* +*ϵ*

Each *𝛽β* value represents the coefficient for the corresponding independent variable (Fernandes, 2024).

**Limitations**

While the model demonstrates significant predictive power, it always assumes the linear relationships and does not account for potential nonlinear interactions between variables. Additionally, the model’s predictive accuracy may be influenced by unobserved confounding variables not included in the dataset.

**B. Ridge Regression**

Ridge regression modifies the least squares objective function by adding a penalty proportional to the square of the magnitude of the coefficients (Hoerl, 1970). The regularization term is controlled by a parameter *𝛼α*.

*Objective=RSS+𝛼∑𝑗=1𝑝𝛽𝑗2*Objective=RSS+*αj*=1∑*p* *βj*2

* **RSS**: Residual Sum of Squares.
* *𝛽𝑗*: Coefficient for feature *𝑗j*.
* *𝛼*: Regularization parameter controlling the impact of the penalty term(Hoerl, 1970).

**C. Lasso Regression Formula**

Lasso regression also modifies the least squares objective but uses the L1 norm of the coefficients, leading to some coefficients potentially being shrunk to zero (Tibshirani, 1996).

*Objective=RSS+𝛼∑𝑗=1𝑝∣𝛽𝑗∣*Objective=RSS+*αj*=1∑*p* ∣*βj* ∣

* + The terms are the same as in Ridge, but the L1 penalty can enforce sparsity in the coefficient vector.

**Alpha**: This is the regularization strength; larger values specify stronger regularization. In Ridge, *𝛼* multiplies the square of the coefficients, whereas in Lasso, *α* multiplies the absolute values of the coefficients.

A grid of *α* values is defined to explore. GridSearchCV systematically works through multiple combinations of parameter values, cross-validating as it goes to determine which tune gives the best performance.

These B, C are foundational in statistical analysis, providing clear insights into the relationships between independent variables and the response variable. They are convenient for their interpretability and simplicity. Regularization techniques (Lasso and Ridge) enhance these models by reducing overfitting and aiding in feature selection, thus improving model robustness.

**D. Random Forest**

This is a type of ensemble tree method that fits a number of decision tree regressors on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control overfitting (Breiman. L, 2001). It creates multiple decision trees during training and outputs the average of their predictions at testing. This process effectively reduces the variance of the predictions, which in turn reduces overfitting.

*Parameters*

* n\_estimators=100: Specifies the number of trees in the forest. Generally, a higher number of trees increases the model's performance and ability to generalize but also increases computational load.
* random\_state=42: Ensures reproducibility of the results by initializing the internal random number generator in a fixed state.

The model is trained using the training data (**X\_train**, **y\_train**). During this phase, the algorithm will construct 100 decision trees, each trained on a random subset of the data points (rows) and features (columns). Each tree in the forest makes its prediction, and the final output is typically the average of these predictions, aiming to provide a robust estimation.

**E. XGBoost**

XGBoost leverages gradient boosting to effectively handle various types of structured data, its ability to manage missing data, and its use of a robust regularization framework to prevent overfitting—features that make it particularly powerful for complex regression tasks (Chen T & Guestrin C, 2016).

**XGBRegressor**

This is the XGBoost class for regression tasks.

* **objective='reg: squared error** tells the model to use the squared error as the loss function during training, which is common for regression problems.
* **random\_state=42** ensures that the results are reproducible with a fixed seed for the random number generator used in XGBoost's bootstrapping.

**Model Evaluation**

During the model evaluation stage in the data science process, machine learning models are assessed for their performance using specific metrics to evaluate their predictive abilities. We utilized a range of metrics, such as RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and R-squared, to compare model performances.

Linear models

Linear regression is used as a baseline model in this analysis. The mean squared error (MSE) is 0.005813 and the R-squared (R²) value is 0.7357. An R² value of 0.7357 means that approximately 73.57% of the variance in the dependent variable is explained by the model. The training and testing R-squared values are almost identical, indicating that the model is neither overfitting nor underfitting. To improve its performance, regularization models like Ridge and Lasso can be considered.

Ridge regression is an improvement over linear regression due to its ability to reduce model complexity and prevent overfitting through L2 regularization. It shrinks the coefficient by penalizing the linear regression. The MSE for Ridge regression is 0.005790, and the R² value is 0.7368, representing a slight increase in R² and a decrease in MSE. This suggests that the model has marginal improvements in fit and predictive accuracy.

Similarly, Lasso regression also incorporates regularization, but it uses L1 regularization that can lead to sparsity in the model. The MSE for Lasso regression is 0.005798 and the R² value is 0.7364, which is close to Ridge's R² value to enhance the fitting by introducing alpha value. This indicates that Lasso also effectively balances bias and variance, but does not substantially outperform Ridge in this scenario.

Ensemble models

The Random Forest model significantly outperforms the linear models with an MSE of 0.002155 and R² value of 0.9020. This indicates a superior predictive performance and a better fit, capable of explaining about 90.20% of the variance. The model benefits from ensemble learning, reducing the risk of overfitting while handling complex nonlinear relationships more effectively.

XGBoost performs comparably to Random Forest, with an MSE of 0.002204 and R² value of 0.8998. Although XGBoost demonstrates slightly higher MSE and marginally lower R², it demonstrates excellent predictive ability and robustness, which are characteristics of gradient-boosting frameworks that iteratively correct errors.

**5. Results**

To comparing results from various models including Linear Regression, Ridge Regression, Lasso Regression, Random Forest, and XGBoost, it's essential to highlight differences in performance using key metrics such as Mean Squared Error (MSE) and R-squared (R²).

**Table 3: Comparision of various models**

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean Squared Error** | **R-squared** |
| Linear Regression | 0.005813 | 0.7357 |
| Ridge Regression | 0.005790 | 0.7368 |
| Lasso Regression | 0.005798 | 0.7364 |
| Random Forest | 0.002155 | 0.9020 |
| XGBoost | 0.002204 | 0.8998 |

Among the linear models (Linear, Ridge, Lasso), Ridge Regression exhibits a slightly better performance with the least Mean Squared Error (MSE) value compared to the other two models. This is due to its regularization effect that helps reduce overfitting. Ridge Regression also shows a marginally higher R-squared value, indicating a better fit to the data than Linear and Lasso Regression.

On the other hand, the Random Forest and XGBoost models show the lowest MSE scores (0.002155 and 0.002204, respectively), indicating that they are more precise in their predictions, with fewer errors on average compared to linear models. Moreover, in terms of R-squared values, with values close to 0.90, which suggests that they can explain about 90% of the variance in the dependent variable. This is significantly higher than the linear models, which hover around 0.735 to 0.737. In summary, evaluating the performance of machine learning models is essential.

The analysis shows that ensemble methods such as Random Forest and XGBoost are better suited for this specific dataset, particularly when dealing with complex patterns that linear models struggle with. These findings can help with the selection of models for similar predictive tasks, particularly in situations where accuracy is essential.

**6. Conclusion**

This study analyzes taxi-in delays at Newark Liberty International Airport (EWR) to improve operational efficiency, reduce costs, and enhance passenger satisfaction. Random Forest and XGBoost models are identified as effective tools for predicting and analyzing taxi-in delays. The findings can suggest areas for targeted operational improvements at EWR, particularly in gate\_management and ground\_handling procedures. By integrating advanced predictive models, EWR can effectively manage and mitigate the impacts of increased traffic, enhancing overall airport efficiency and passenger experience. These findings provide a clear direction for future research and operational strategies at EWR and other airports facing similar challenges. So, we can conclude that taxi-in delay is not that contributing factor as taxi-out delay.

**ACKNOWLEDGMENT**

"We are grateful to Dr. Tony Diana for his exceptional guidance and unwavering support throughout this project. His continuous motivation and valuable feedback have been instrumental in refining our ideas and enhancing the quality of our work. I deeply appreciate his leadership and the time he invested in assisting me."

**7. References**

Diana, T. (2018). Can machines learn how to forecast tax-out time? A comparison of predictive models applied to the Seattle/Tacoma International Airport case.

Anupkumar, A. (2023). Investigating the costs and economic impact of flight delays in the aviation industry and the potential strategies for reduction. Electronic Theses, Projects, and Dissertations.

Choi, S., & Kim, Y. J. (2021). Artificial neural network models for airport capacity prediction. Journal of Air Transport Management.

Dinler, N., & Rankin, W. B. (2020). Increasing airport easing airports' on-time arrival performance through capacity and efficiency indicators. International Journal of Aviation, Aeronautics, and Aerospace.

Park, D. K., & Kim, J. K. (2018). Influential factors to aircraft taxi time in the airport.

Zhang, L., & Wang, Y. (2010). Statistical modeling of aircraft taxi times: Insights from a comparative study. Journal of Air Transport Management.

Johnson, D. E., Thompson, K., & Gupta, S. (2017). Analyzing the impact of air traffic congestion on taxi-in delays. Transportation Research Part A: Policy and Practice.

Lee, M. K., & Park, D. K. (2021). Optimizing gate assignments to minimize taxi-in delays at busy airports. Journal of Operations Management.

Park, D. K. (2022). A comprehensive review of taxi-in delay dynamics and predictive analytics in aviation. Progress in Aerospace Sciences.

Géron, A. (2017). Hands-On Machine Learning with SciKit-Learn & TensorFlow. O’Reilly, Sebastopol, CA.

Hebert, J. E., & Dietz, D. C. (1997). Modeling and analysis of an airport departure process.

Journal of Aircraft.

Fernandes, V., Carvalho, G., Pereira, V., Bernardino, J., & Bernardino, J. (2024). Analyzing Data Reduction Techniques: An Experimental Perspective. Applied Sciences

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal*  *Statistical Society. Series B (Methodological), 58*(1).

Breiman, L. (2001)**.** "Random Forests." *Machine Learning, 45*(1), 5-32. This seminal paper by Leo Breiman introduces the Random Forest algorithm and discusses its statistical properties.

Chen, T., & Guestrin, C. (2016)**.** "XGBoost: A Scalable Tree Boosting System." In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, New York, NY, USA.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data*  *Mining, Inference, and Prediction* (2nd ed.). Springer Series in Statistics.