

# EXPERIMENT REPORT

Student Name	Panalee Makha
Project Name	ML Data Product
Date	9 November 2023
Deliverables	Makha_Panalee-14367914-knn.ipynb Linear Regression, Ridge, ElasticNetCV, K-Nearest Neighbors, Gradient boosting <a href="https://github.com/wongwara/flight-streamlit-at3/blob/main/flight-prediction/notebooks/PM_notebooks/model_regression-2.ipynb">https://github.com/wongwara/flight-streamlit-at3/blob/main/flight-prediction/notebooks/PM_notebooks/model_regression-2.ipynb</a>

## 1. EXPERIMENT BACKGROUND

Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach.

### 1.a. Business Objective

The objective of this project is to create models for the development of a data product aimed at assisting users in the USA in improving their accuracy when estimating local travel airfare. Users will have the capability to input their trip specifics, and the application will provide predictions for their expected flight fares.

### 1.b. Hypothesis

The Linear Regression, Ridge, ElasticNetCV, Gradient boosting, and K-Nearest Neighbors models are expected to demonstrate a low level of mse and mae when predicting the airfare for a particular trip.

Perform a comprehensive evaluation of all the models in order to determine the model that results in the highest level of accuracy in its predictions.

### 1.c. Experiment Objective

The key objective of this endeavour is to accurate Airfare Estimation by developing a data product that can accurately estimate local travel airfare for users in the USA, providing reliable fare predictions. This requires data collection, feature engineering, model training with Linear Regression, Ridge, ElasticNetCV, Gradient boosting, and K-Nearest Neighbors and comparative analysis to determine the most accurate prediction. The research aims to improve airfare predicting decision-making, which will ultimately contribute to enhanced business planning and performance.

## 2. EXPERIMENT DETAILS

Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them.

### 2.a. Data Preparation

- **Data preprocessing**
  - **Merging Datasets:** Merge datasets from various airports into a single dataframe for exhaustive analysis.
  - **Converting Travel Duration:** Convert travel duration from a string format into seconds for consistency and easy analysis.
  - **Segment Splitting:** The segment columns are split into lists.
  - **Total Segment Duration:** Calculate the total duration of segments and create a new column named "totalDuration (segment)."
  - **Travel Layover:** A new column, "travelLayover," represents the difference between travel duration and the total segment duration.
  - **Transit Airport Code:** Extract and store the airport code of transit locations in a new column named "transitAirportcode."
  - **All Airport Codes:** We collect all airport codes, including departure, arrival, and transit, and store them in a new column labelled "All airport."
  - **Date Transformation:** The date column serachDate and flightDate is transformed into separate columns for day, month, and year.
  - **Departure Time Extraction:** Extract the departure time in Unix epoch seconds format and separate it into individual columns for hours, minutes, and seconds.
  - **Handling Null Values:** To address missing data, fill null values with the mean of the respective columns.
  - **Feature selection:** The selected features, 'totalTravelDistance,' 'isNonStop,' 'isBasicEconomy,' 'startingAirport,' 'destinationAirport,' 'segmentsCabinCode,' 'flightDate\_day,' 'flightDate\_month,' 'flightDate\_year,' 'DepartTime\_hour,' 'DepartTime\_minute,' 'DepartTime\_second,' and 'totalFare.'
- **Data splitting:** The dataset was divided into training, testing, and validation sets with an 80:20 ratio. It is important to note that all the experiments used the same dataset after this split, allowing for a fair comparison of the best model later.

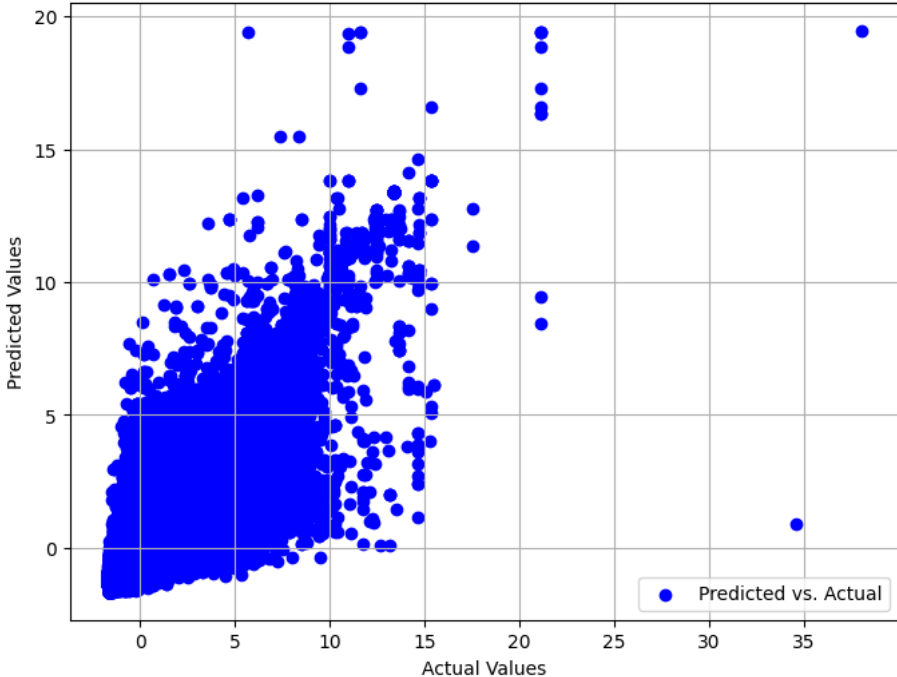
<p>2.b. Feature Engineering</p>	<ul style="list-style-type: none"> <li>• <b>Feature Engineering</b> <ul style="list-style-type: none"> <li>▪ <b>Factorising Segment Columns:</b> The segment columns were factorised to enhance their usability in modelling.</li> <li>▪ <b>Label Encoding 'SegmentCabinCode':</b> Applied label encoding to the 'segmentCabinCode' column, converting categorical data into a numerical format for analysis.</li> <li>▪ <b>Encoding Categorical Columns:</b> Other categorical columns were also encoded using label encoding, ensuring consistency in handling categorical data.</li> <li>▪ <b>Scaling Numeric Columns:</b> Numeric columns were scaled using Standard Scaler, making them compatible with the modelling process.</li> </ul> </li> </ul>
<p>2.c. Modelling</p>	<p><b>Linear Regression, Ridge, ElasticNetCV, Gradient boosting, K-Nearest Neighbors</b> : These hyperparameters were chosen based on common practices and considerations for KNN models</p> <ul style="list-style-type: none"> <li>• neighbours: [ 3, 5, 7, 9 ] - The range of neighbours provides flexibility in capturing local patterns</li> <li>• weights: [ 'uniform' ] - Treat all neighbours equally</li> <li>• p: [ 1 ] - the Manhattan distance metric</li> </ul> <p>The MSE and MAE measures will be used to evaluate model performance.</p>

### 3. EXPERIMENT RESULTS

Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified.

#### 3.a. Technical Performance

MSE		
Model	Train set	Validation set
Linear regression	0.5797947093319961	0.5758296803478751
Ridge alpha=10.0	0.5797947093498554	0.5758296777111511
ElasticNetCV	0.5799352635870615	0.5759680777159685
KNN Default	0.21086204154450128	0.2886026235149159
KNN n_neighbors=3, weights='uniform', p=1	0.22689346780849887	0.2472002908678711
KNN n_neighbors=5, weights='uniform', p=1	0.28851670394012663	0.22908974023913908
KNN n_neighbors=7, weights='uniform', p=1	0.20673803854627698	0.2237905743167268
KNN n_neighbors=9, weights='uniform', p=1	0.20617295706992672	0.22214024746891006
Gradient Boosting	0.4366272059715127	0.46549381168591236
MAE		
Model	Train set	Validation set
Linear regression	0.5339013787088485	0.5334993698926779
Ridge alpha=10.0	0.5339012467864863	0.5334992382406365
ElasticNetCV	0.533719626884408	0.5333240964458702
KNN Default	0.22950249114460528	0.2992897143886497
KNN n_neighbors=3, weights='uniform', p=1	0.2985561514379969	0.3103953942965042
KNN n_neighbors=5, weights='uniform', p=1	0.2106519096406004	0.2991453010843101
KNN n_neighbors=7, weights='uniform', p=1	0.2854750925942013	0.29528471956271546
KNN n_neighbors=9, weights='uniform', p=1	0.284524113772312	0.2937953365697721
Gradient Boosting	0.435358341010421	0.46520639055619195

	<div>Best model on Test set</div> <table><thead><tr><th>Model</th><th>MSE</th><th>MAE</th></tr></thead><tbody><tr><td>KNN n_neighbors=9, weights='uniform', p=1</td><td>0.22154418857678332</td><td>0.29403444429653136</td></tr></tbody></table> <div><p>Predicted vs. Actual Values</p></div>	Model	MSE	MAE	KNN n_neighbors=9, weights='uniform', p=1	0.22154418857678332	0.29403444429653136
Model	MSE	MAE					
KNN n_neighbors=9, weights='uniform', p=1	0.22154418857678332	0.29403444429653136					
3.b. Business Impact	<p>The results of training five different regression models with notable performance distinctions. The top-performing model is KNN4 or K-Nearest Neighbors with neighbours=9, weights='uniform', p=1, showcasing the lowest MSE on both the training and validation sets, approximately 0.206 and 0.222, respectively. Following closely is the Gradient Boost model, exhibiting slightly superior MSE scores compared to Linear regression, Ridge regression, and the ElasticNetCV model on both training and validation sets.</p> <p>The models offer users substantial advantages, delivering precise and dependable airfare estimates for travel planning. By addressing challenges in budget planning and fare analysis, the model enables more informed travel decisions, aligning with users' needs and enhancing their overall experience.</p>						
3.c. Encountered Issues	<ul style="list-style-type: none"><li>• Due to the large file size of the KNN model files, knn_fit.joblib and knn_fit.joblib.zip, they cannot be uploaded to GitHub. The best KNN model, which is saved in the models folder, does not include fitted data which can access the KNN model file, including the fitted data, through the following Google Drive link: <a href="https://drive.google.com/drive/folders/12ISspcn9g2bXIPBGeUGjVq1uTjT2XJup">https://drive.google.com/drive/folders/12ISspcn9g2bXIPBGeUGjVq1uTjT2XJup</a></li><li>• Due to prolonged runtime exceeding 200 minutes while attempting GridSearchCV for parameter optimization in Gradient Boosting and KNN models, the approach was abandoned. Instead, hyperparameter tuning was adopted for efficiency.</li></ul>						

#### 4. FUTURE EXPERIMENT

Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective.

##### 4.a. Key Learning

The KNN models demonstrate superior performance compared to linear regression models, showcasing notably reduced MSE and MAE values. It's interesting to note that with an increase in the number of neighbors (K) from 3 to 9, training MSE and MAE values decrease, indicating enhanced model fit. However, the diminishing improvement in validation MSE and MAE values suggests a potential risk of over-smoothing with higher K values.

##### 4.b. Suggestions / Recommendations

The precision of fare predictions is intricately tied to the quality of the underlying data. To further enhance prediction accuracy, consider implementing a strategy of continuous dataset updates and refinement of data sources. Regularly incorporating the latest information into the model can lead to more accurate and reliable predictions. Additionally, exploring diverse and comprehensive data sources could provide a more nuanced understanding of factors influencing airfare, ultimately contributing to more informed and precise predictions.