Project Report

|  |  |
| --- | --- |
| **Course Name (NICF)** | Data Science Essentials |
| Product Name (Marketing & Sales) | ACDS |
| **Module Name (NICF)** | Data Science Essentials |
| Product Name (Marketing & Sales) | ACDS |

|  |  |  |  |
| --- | --- | --- | --- |
| Student name | | Assessor name | |
| Tan Tiong Choon (Chen ZhongChun) | | Ryan Suryanto | |
| Date issued | Completion date | | Submitted on |
| 14 Nov 2023 | 23 Nov 2023 | | 23 Nov 2023 |
|  | |  | |
| Project title | |  | | --- | | Train, Test and Publish a Regression Model using Azure Machine Learning Platform | | | |

|  |
| --- |
| Learner declaration |
| I certify that the work submitted for this assignment is my own and research sources are fully acknowledged.    Student signature: Date: 23 Nov 2023 |
|  |

**1. Project Overview**: Describe the Project with summary of analytical processes and project outcomes (Explain the Project in your own words in 15 – 20 lines)

This project requires us to implement a predicting machine learning model using Microsoft Azure Machine Learning platform with the dataset provided by the Azure Machine Learning platform. It provides us with the flight departure and arrival data so we can understand the critical features that might be predictive of how many minutes late or early a flight will be. It also enables us to have an overall view of the machine learning environment on how this predictive modelling is build, operate and function.

We will need to verify and analyse data related to flight departure and arrival information, these will enable us to identify the trends, patterns and correlations over time. We are using labelled data to train the model using regression method to predict the arrival delay timing. We will use the necessary models to remove duplication, handle missing values and normalise the data to prepare the model for evaluation and publishing for web service. The result can be access online and test and evaluate by using Microsoft Excel by inputting the necessary information and predicted label will be shown.

The expected outcomes will be train, evaluate and publish a model showcasing flight arrival delay timing, with a focus in accuracy and reliability. It gives us insights into flight punctuality and enable us to identify and analysis of predictive features that will affect the predicted result. We shall be presenting a predictive model to predict the arrival delay time that will assist the stakeholders, policymakers and researchers to understand the scale and impact of the delay. With this forecast, we shall be able to present the audience with better understanding of the delay involved and identify the potential area for proactive measure to be taken.

**2. Project Technical Environment: (Describe the Microsoft Azure Machine Learning Platform)**

We shall be using Microsoft Machine Learning Studio for building, training and deploying machine learning models.

The data will be provided by using Microsoft Machine Learning Studio which contains 1 dataset file containing the flight details.

We can upload and manage flight dataset which include year, month, carrier, departure or arrival time, departure or arrival delay time etc.

Microsoft Machine Learning Studio is a cloud-based platform that do not required any programming unless necessary for Python/R script. It allows the data scientist to focus on developing the experiment using drag and drop tools. It allows us to remove duplicated rows, cleaning missing data, make categorical data, normalize data, split data, add regression model, train, score and evaluate the model.

We shall use Boosted Decision Tree Regression to train the model for ArrDelay and publish and use the model.

3. Analytical Technique & Tools used: Describe the Analytical Technique and Tools used in the Project

The integrated development environment we will be using is Microsoft Machine Learning Studio.

We will be using Remove Duplicated Rows, Cleaning Missing Data, Edit Metadata, Normalize Data, Split Data, Boosted Decision Tree Regression Model to train, score and evaluate this experiment and thereafter publish and use the model.

We will be using Visualize of each model to explore the flight dataset to understand its characteristics and identify patterns or anomalies.

We will be transforming the data using Cleaning Missing Data, Edit Metadata, Normalize Data model make it more suitable to modelling.

4. Data Science Project Team – Roles and Responsibilities Table

Project Manager: Define project objective, budget and scope, develop timelines, allocating and coordinating resources.

Data Scientist: Develop machine learning model for predictive analysis, explore various algorithms, explore and pre-process the data and collaborate with domain expert.

Data Engineer: Data quality and consistency maintenance, data extraction and transformation, database management.

Business Analyst: Understand business goals and objectives, customer requirements, communicate insights to stakeholders, bridges gaps between technical teams and business stakeholders.

Data Analyst: Collect and compile data from various sources, exploratory data analysis, create data visualization and reports.

Domain Expert: Provide domain specific knowledge and contents.

5. Activity 1: Activity Summary

Task 1: List and explain the roles of a Data Science team

Project Manager:

* Define project scope, objectives, scopes, budget.
* Develop project plans and schedules.
* Allocate resource and manage budget.
* Communicate with stakeholders on schedule and requirements.
* Monitor and report project progress.
* Resolve issues.

Data Scientist:

* Develop and apply machine learning models and algorithms.
* Clean, preprocess, and analyze data to identify patterns and trends.
* Interpret and communicate findings or insights and make recommendations to business leaders
* Collaborate with domain experts to understand business objectives
* Develop feasible data-based recommendations that will address organisational problems.
* Stay updated with the latest advancements in data science, machine learning, and relevant tools and technologies.

**Data Engineer:**

* Design and maintain data pipelines to collect, process, and store data.
* Ensure data quality and integrity.
* Optimize data storage and retrieval for analysis.
* Optimize data workflows for efficiency and performance.
* Collaborate with data scientists and analysts to understand their data requirements.

Business Analyst:

* Work closely with stakeholders to understand business objectives and requirements.
* Conduct interviews, workshops, and surveys to gather information from business users.
* Use data to support decision-making and help identify areas for process improvement.
* Facilitate communication between business stakeholders and the technical team.
* Verify that implemented solutions align with the initially defined objectives.
* Generate reports for management and stakeholders.

**Data Analyst:**

* Collect and compile data from various sources
* Conduct exploratory data analysis.
* Create data visualizations and dashboards to communicate insights.
* Assist in data interpretation and reporting.
* Support data scientists in data preparation and analysis.

Domain Expert:

* Provide subject matter expertise in the relevant domain.
* Define and clarify business requirements.
* Collaborate with data scientists to validate models and interpretations.

Task 2: Review the Data Cleansing, Exploration and Machine learning activities.

Data Cleaning: Handle Duplicated Data, Handle Missing Data, Handle Outliers. The primary objectives of data cleaning are to enhance the quality of the data, ensure its accuracy, and create a reliable analysis or modelling.

Data Exploratory: Investigate, identify patterns, understand the dataset. The primary objectives of data exploratory is to maximise insights into a dataset and uncover patterns, trends, anomalies, and relationships before moving on to more advanced analyses or modelling.

Machine Learning: Transform, normalize, split data so that we can split, train and evaluate the data for production environment. The primary objectives of machine learning is to create intelligent systems that can learn from data, adapt to changing conditions, and provide valuable insights for decision-making.

Task 3: Prepare a report of 1-2 pages explaining the output of each activity in 2-3 lines for each activity

Data Cleaning:

* Identify and analyse missing values in the dataset.
* Decide on how to handle missing data, whether to delete, replace values with mean or zero.
* Identify and remove duplicate records or rows from the dataset.
* Standardize data formats to ensure consistency.
* Identify outliers or extreme values that may be errors or anomalies.
* Standardize categories in categorical variables.

Data Exploratory:

* Computing and summarizing mean, median, mode, minimum, maximum, standard deviation to understand the data.
* Creating visual representations of the data using charts, graphs, histograms and scatter plots. It helps in identifying patterns and outliers, and it provides an understanding of the data distribution
* Explore relationships between pairs of variables (bivariate analysis) and assess interactions between multiple variables (multivariate analysis).
* Calculate correlation coefficients to quantify the strength and direction of relationships between numerical variables.
* Creating new variables based on existing ones to enhance the modelling.
* Identify and address missing values, outliers, and inconsistencies.

Machine learning:

* Clearly articulate the problem or task you want the machine learning model to solve. Define the goals, objectives, and success criteria.
* Gather relevant data that will be used to train and evaluate the machine learning model. Ensure the data is representative of the problem you are trying to solve.
* Clean the data by addressing missing values, duplicated data, handling outliers, and ensuring consistency.
* Clean missing data by selecting column names, you may specify the minimum and maximum ratio. Specify the mode that you wish to use for cleaning such as mean, median, mode, remove row or column, substitute value.
* Pre-process the data to make it suitable for model training, which may involve scaling by prevents certain features with larger scales from dominating, encoding categorical variables which will enhance their understanding of the data, facilitating better analysis and interpretation.
* Divide the dataset into training and test sets. The training set is used to train the model, and the test set is used for evaluating the model's performance.
* Select the type regression or classification for training the model. Classify data into distinct categories or predict numerical values. Classification and regression are fundamental tasks in machine learning that address a wide range of applications
* Train the selected model using the training dataset. The model learns from the input features and their corresponding target values.
* Use the score model to predict values. Evaluation the model from score model will show the metrics such as root mean squared error, mean/relative absolute error, relative squared error, coefficient of determination.
* Deploy the trained model into a production environment if the performance meets the desired criteria.
* Iterate on the model and the entire machine learning pipeline based on feedback, new data, or changes in the problem domain. Continuous improvement is a key aspect of machine learning system development.

Activity 2: Flight Dataset- An Overview

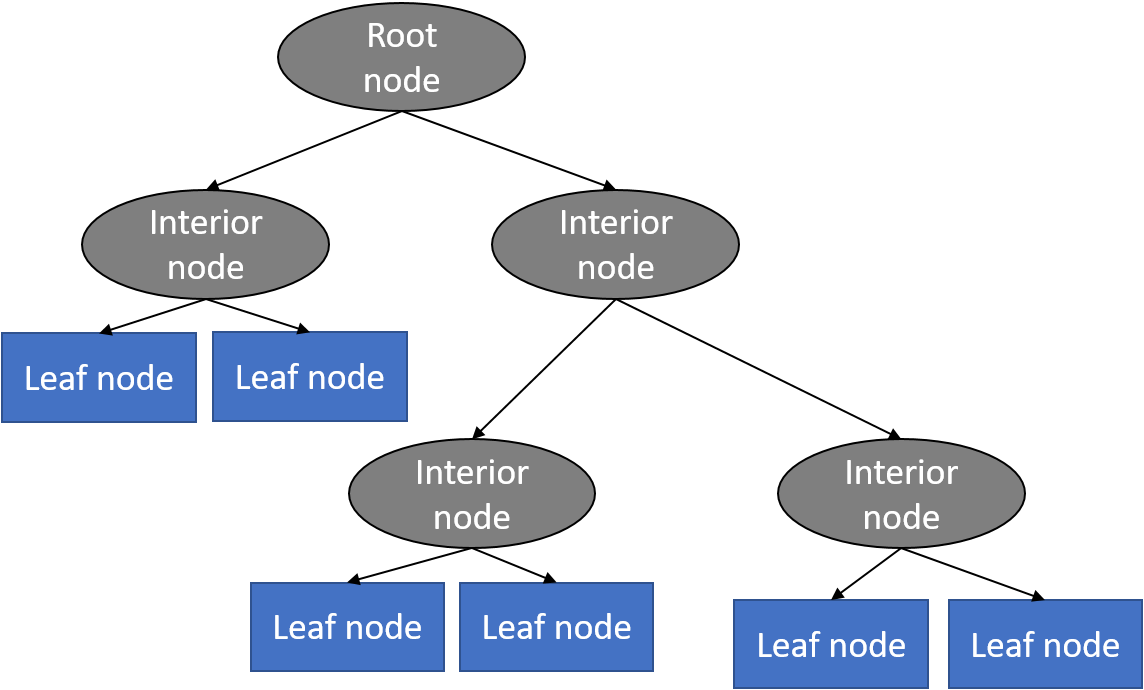
Task 3: Explain the experiment and Regression/ Clustering model used in a single Page and include it in the Project Report

The experiment is about implement a predicting machine learning model to predict arrival delay time. It enables us identify the trends, patterns and correlations over time. We are training the experiment using regression method to predict the arrival delay timing. We will be cleaning the data using the Remove Duplicated rows, Cleaning Missing Data models. Pre-processing data using Edit Metadata and Normalise data model. Splitting the data into 70% for training and 30% for testing using Split data Module. We will be training the model and using Boosted Decision Tree Regression model and Score Model and Evaluate Model to prepare the model for evaluation and publishing for web service. The result can be test and evaluate by using Microsoft Excel by inputting the necessary information and predicted label will be shown.

A Decision Tree Regression model is a type of supervised learning algorithm used for predicting continuous numeric values. Decision tree builds regression or classification models in the form of a tree structure. While Decision Trees are often associated with classification tasks where the goal is to assign data points to distinct categories or classes, they can also be adapted for regression tasks

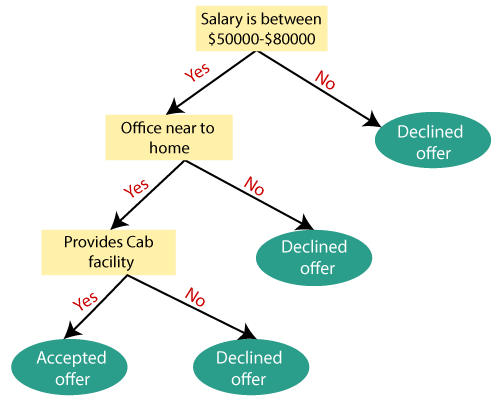
It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

It is a tree-structured classifier with three types of nodes. The ***Root Node***is the initial node which represents the entire sample and may get split further into further nodes. The ***Interior Nodes***represent the features of a data set and the branches represent the decision rules. Finally, the ***Leaf Nodes*** represent the outcome. This algorithm is very useful for solving decision-related problems.



With a particular data point, it is run completely through the entirely tree by answering *True/False* questions till it reaches the leaf node. The final prediction is the average of the value of the dependent variable in that particular leaf node. Through multiple iterations, the Tree is able to predict a proper value for the data point.

The calculation is consisted of counting the number of items, the average, standard deviation and coefficients of variation.



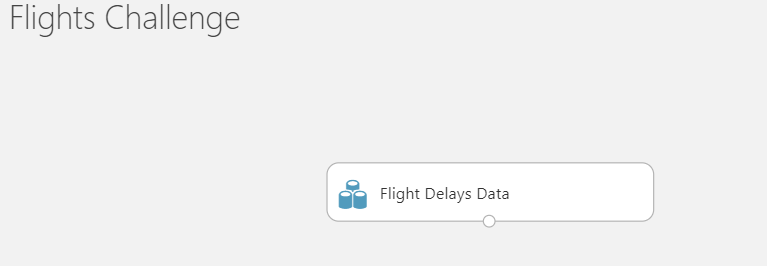
The above diagram is a representation for the implementation of a Decision Tree algorithm. Decision trees have an advantage that it is easy to understand, lesser data cleaning is required, non-linearity does not affect the model’s performance and the number of hyper-parameters to be tuned is almost null.

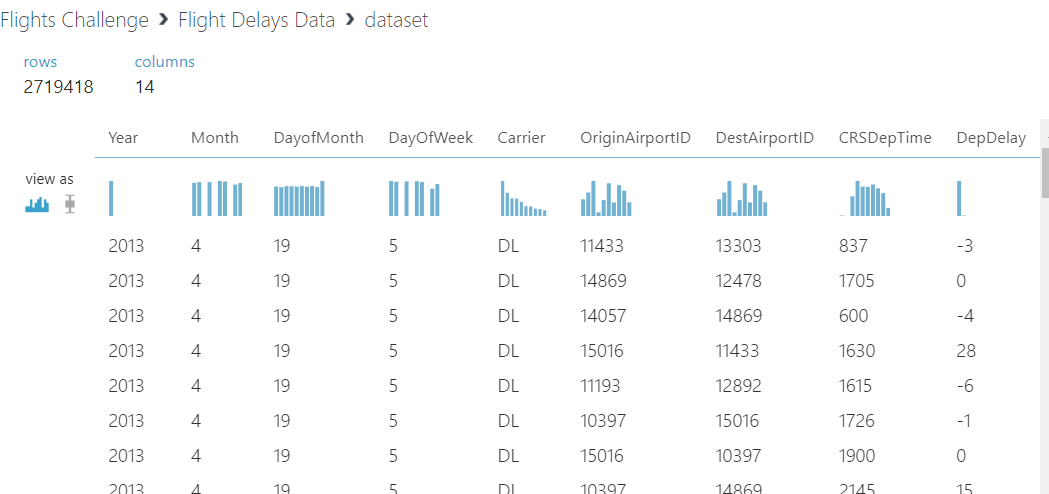
The construction of the tree continues until a predefined stopping criterion is met. This could be a maximum depth for the tree, a minimum number of data points in a leaf node, or other conditions.

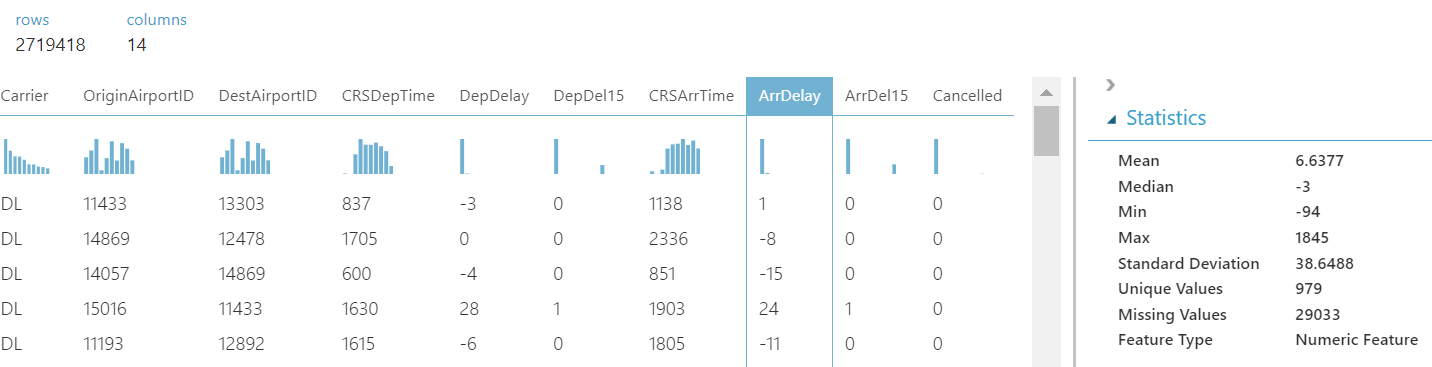
Activity 3: Create New Experiment with existing data storage mechanism.

*Task 2:* Create a new experiment, with an appropriate name like "Flights Challenge".

*Task 3:* Add the Flights Delay Data sample dataset to the experiment, and then visualize its contents.







Task 4: Answer the following questions:

How many rows are in the dataset?

- What is the **mean** value of the **ArrDelay** column?

There are 2719418 rows and the mean value of ArrDelay is 6.6377

Activity 4: Remove Duplicates and Replace Missing Values [ ETL process)

*Task 1:* Remove duplicate rows (retaining the first instance of each row). Rows are considered duplicates in this dataset if they have matching values for all the following fields:

- Year

- Month

- DayofMonth

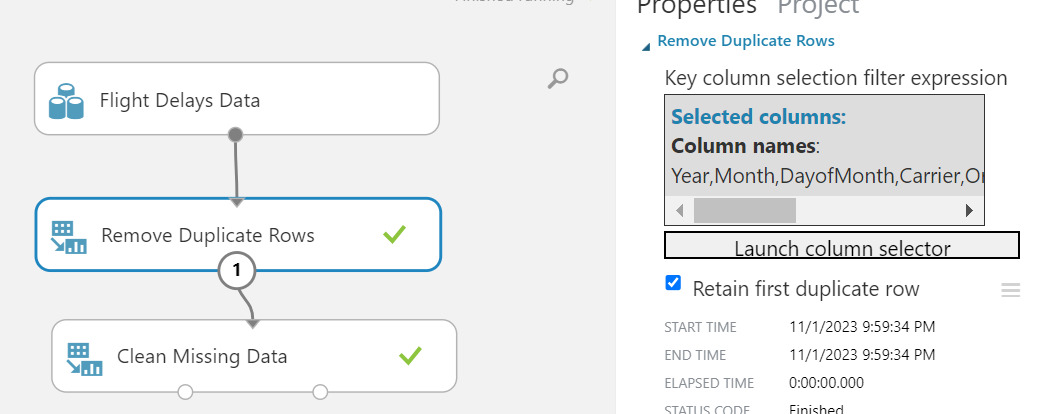
- Carrier

- Origin Airport ID

- Dest Airport ID

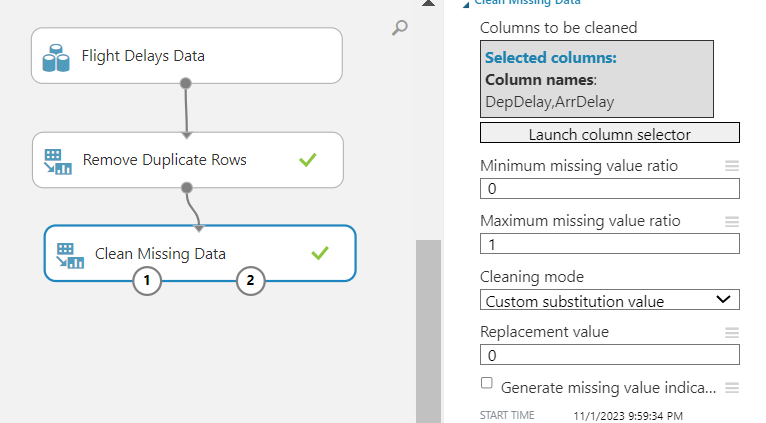
- CRS Dep Time

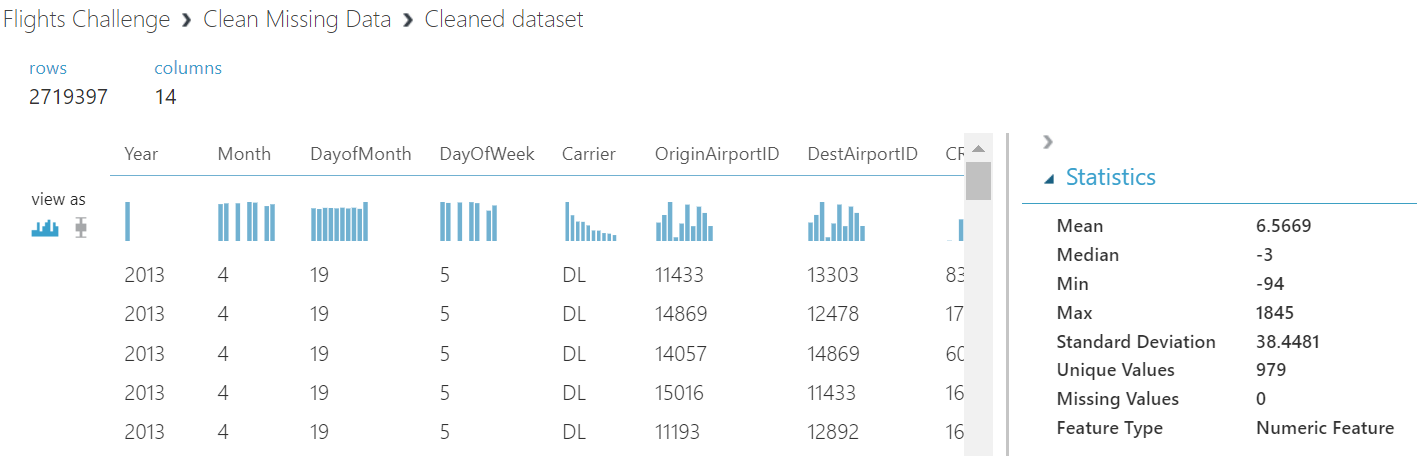
- CRR Arr Time





Task 2: After removing the duplicate rows, replace missing values in the DepDelay and ArrDelay columns with the value 0 (zero). Use the built-in Azure Machine Learning Cleaning Missing Data module,





**Task 3:** Answer the following questions, after you have removed duplicate rows and replaced missing values,

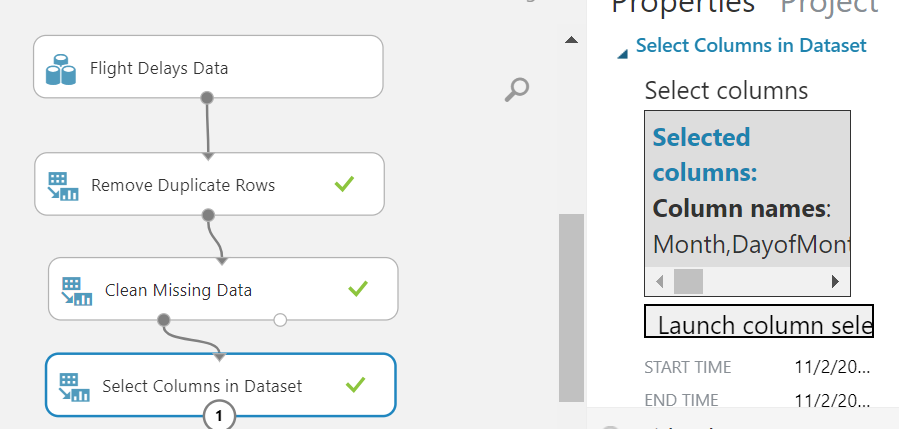
- How many rows remain in the dataset?

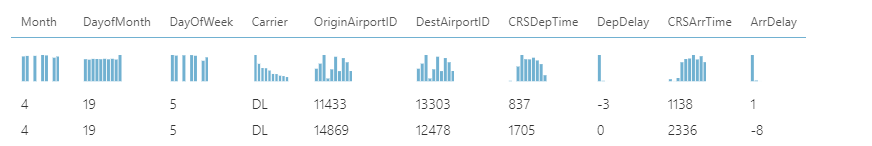
- What is the mean value of the ArrDelay column?

There are 2719397 rows and the mean value of ArrDelay is 6.5669

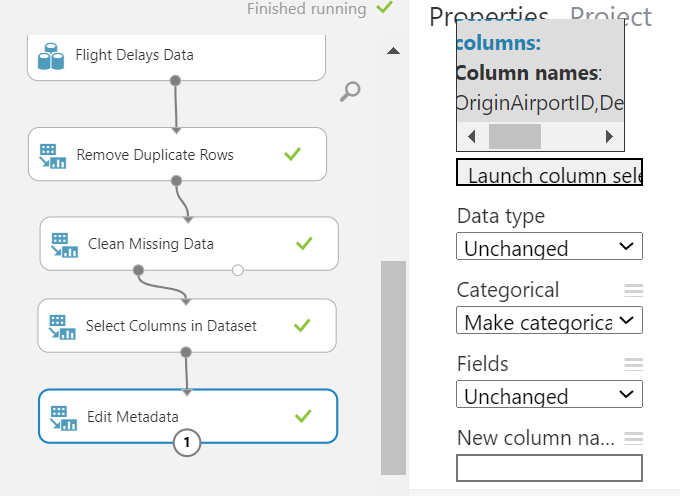
Activity 5 : Train a Regression Model

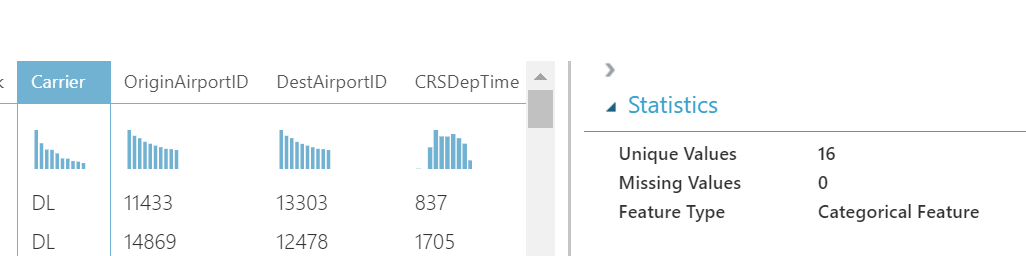
Task 2: Add a **Select Columns in Dataset** module, and use it to select only the **Month, DayofMonth, DayOfWeek, Carrier, OriginAirportID, DestAirportID, CRSDepTime, DepDelay, CRSArrTime**, and **ArrDelay** columns.

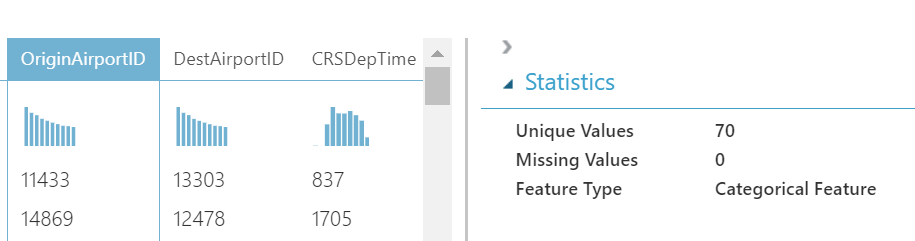


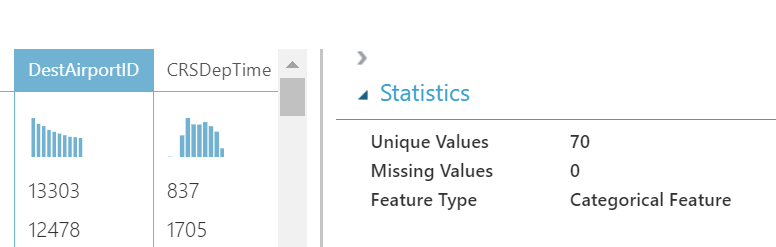


Task 3: Add an **Edit Metadata** module and use it to make the **OriginAirportID, DestAirportID,** and **Carrier** columns **Categorical**.

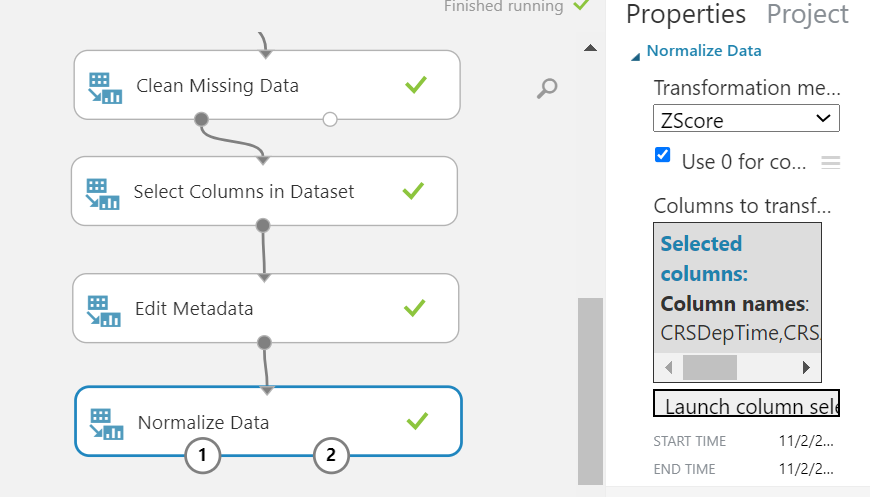


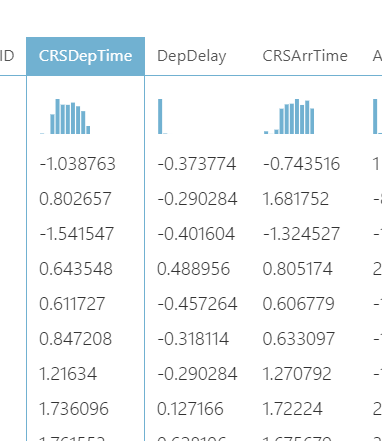




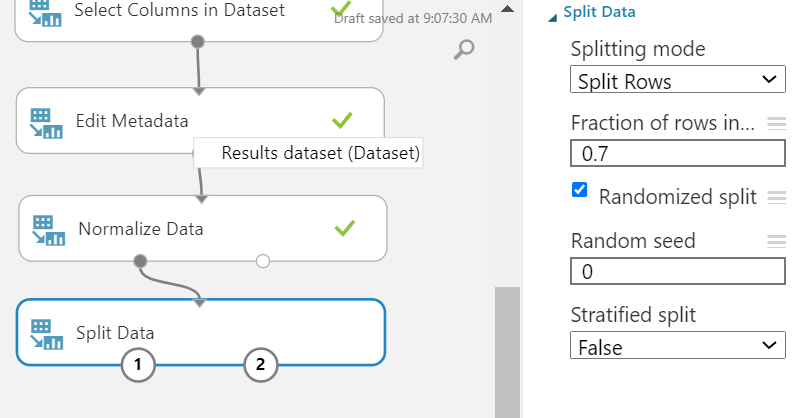


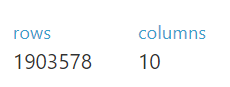
Task 4: Add a **Normalize Data** module and use it to standardize the **CRSDepTime, CRSArrTime,** and **DepDelay** columns using the **ZScore** transformation method.



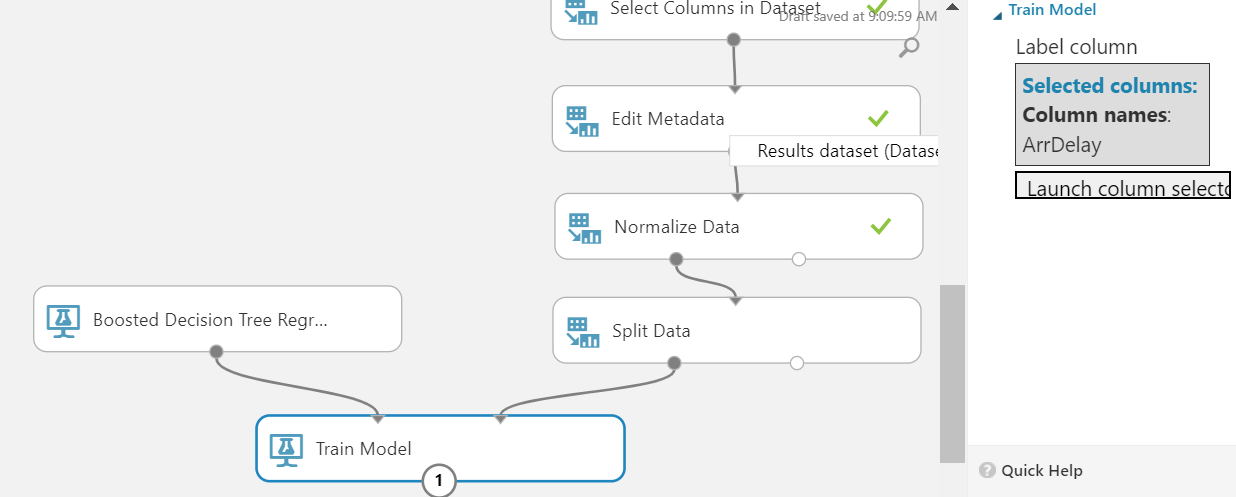


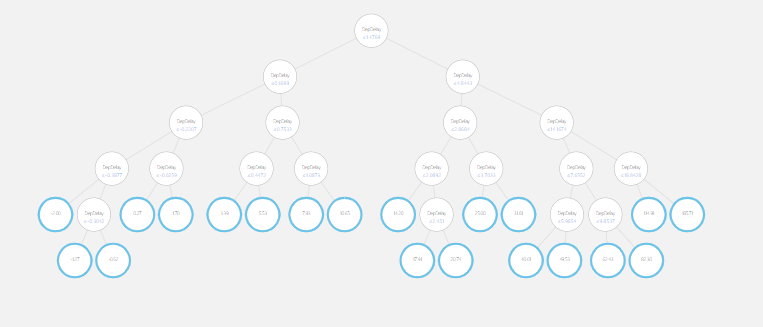
Task 5: Add a **Split Data** module and use it to split the rows into 70% / 30% subsets. Use a random seed value of **0**.



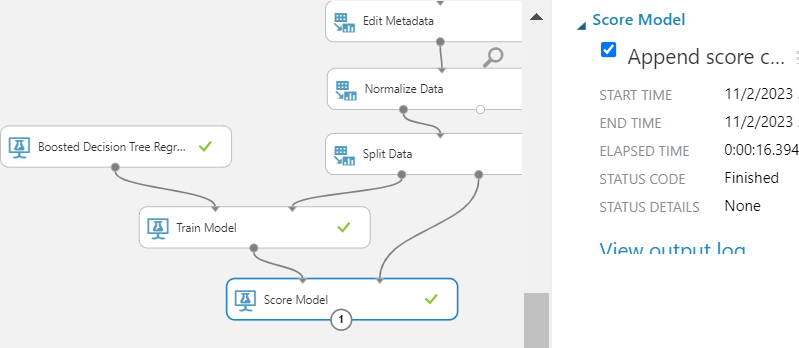


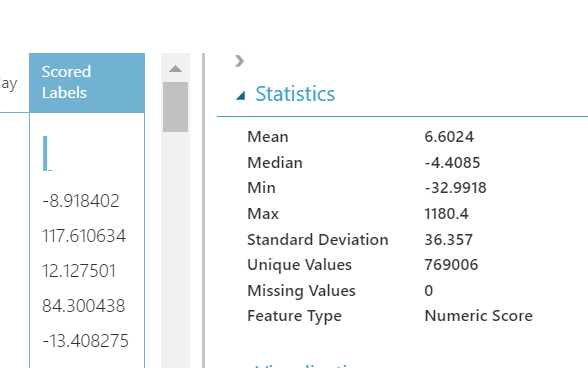
Task 6: Add a **Boosted Decision Tree Regression** module and a **Train Model** module. Then use the default settings to train the model with the 70% data split to predict the **ArrDelay** label column.



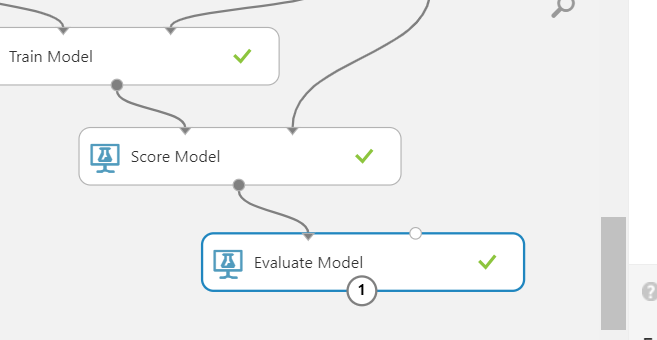


Task 7: Add a **Score Model** module, and use it to score the trained model using the 30% split of data.





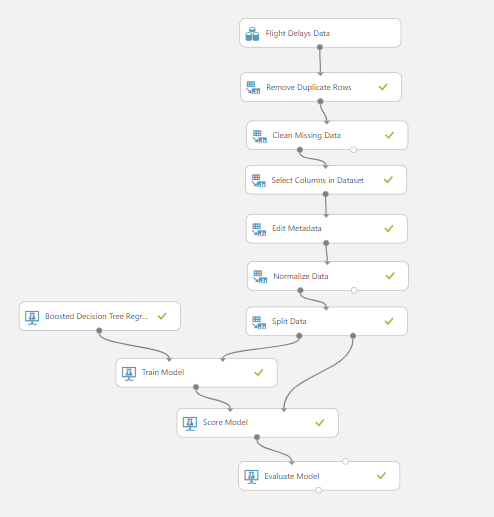
Task 8: Add an **Evaluate Model** module and use it to evaluate the results from the **Score Model** module.

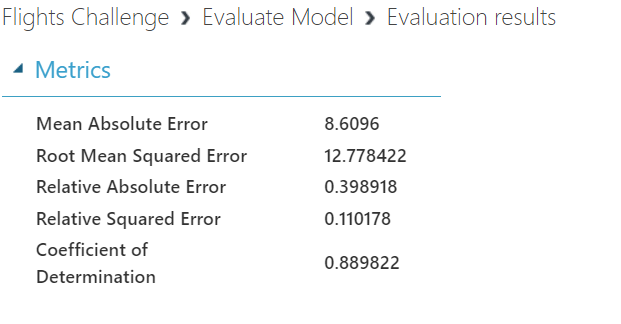


Activity 6: Test and Evaluate the Model

Task 1: Run the experiment,

Task 2: When it has finished, visualize the output of the Evaluate Model module.



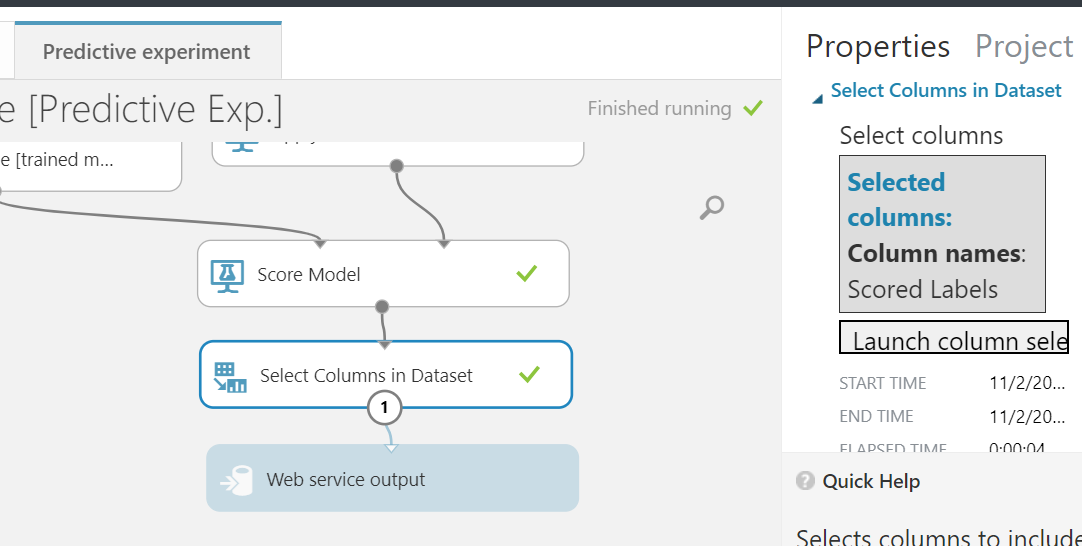


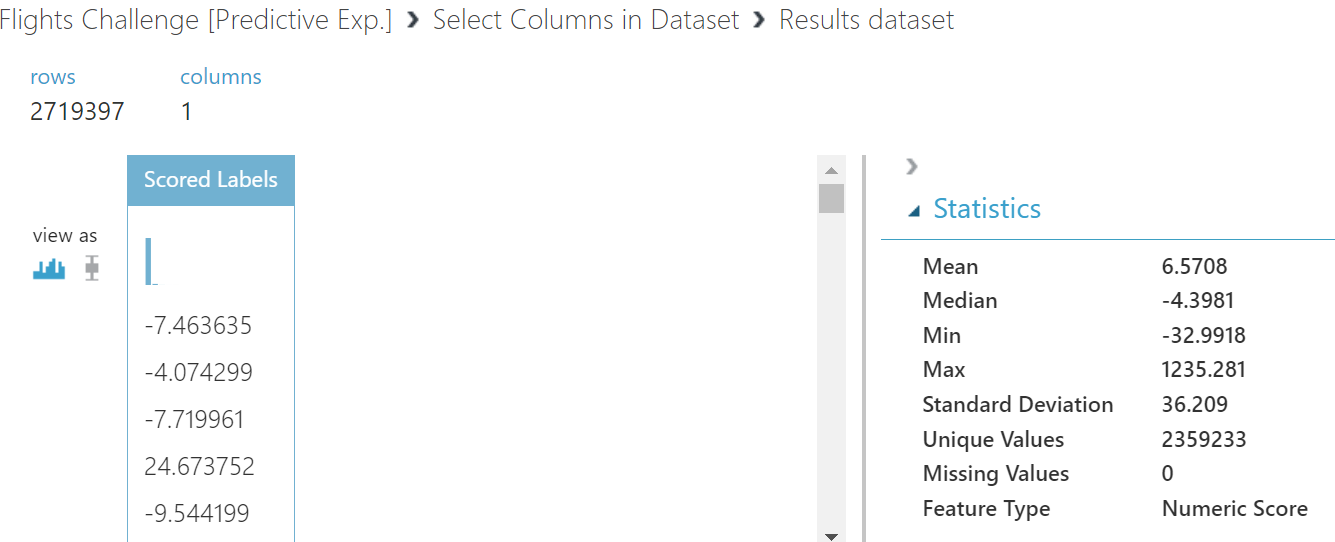
Root mean square error is to evaluate the accuracy of the predictive model. A value of 12.77 which is considered not so good as value closer to 0 indicate smaller prediction error between predicted and actual values.

Coefficient of Determination is a measure of how well observed outcomes are replicated by the model. The value of 0.88, is considered a good result as value closer to 1, the model perfectly explains the variability in the dependent variable.

Activity 7: Publish and Use the Model

*Task 2:* In the predictive experiment, add a Select Columns in Dataset module and place it between the Score Model and Web service output modules. Use this to select only the Scored Labels column.





*Task 3:* Save and run the modified predictive experiment, and then deploy the web service.

