Project Report

|  |  |
| --- | --- |
| **Course Name (NICF)** | **NICF Diploma in Infocomm Technology (Data)** |
| Product Name (Marketing & Sales) | Professional Diploma in Data Science |
| **Module Name (NICF)** | **NICF Data Science Essentials (SF)** |
| Product Name (Marketing & Sales) | Data Science Essentials |

|  |  |  |  |
| --- | --- | --- | --- |
| Student name | | Assessor name | |
| Lee Jack Shiang | | Ryan Suryanto | |
| Date issued | Completion date | | Submitted on |
| 9-Jun-2022 | 15-Jun-2022 | | 15-Jun-2022 |
|  | |  | |
| 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: 15-Jun-2022 |

Content

1. Project Overview 3
2. Project Technical Environment 4
3. Analytical Technique & Tools used 5
4. Data Science Project Team – Roles and Responsibilities Table 6
5. Activity 1: Activity Summary 7
6. Activity 2: Flight DataSet- An Overview 9
7. Screen-shots of each task of Activity 3 10
8. Screen-shots of each task of Activity 4 11
9. Screen-shots of each task of Activity 5 12
10. Screen-shots of each task of Activity 6 14
11. Screen-shots of each task of Activity 7 16
12. Screen-shots of each task of Activity 8 19
13. Screen-shots of each task of Activity 9 26
14. Screen-shots of each task of Activity 10 31
15. Screen-shots of each task of Activity 11 36
16. Screen-shots of each task of Activity 12 38
17. Conclusion 40
18. Annexure – 01 41
19. Annexure – 02 42

Project Overview: Describe the Project with summary of analytical processes and project outcomes (Explain the Project in your own words in 15 – 20 lines)  
  
In this project, Azure Machine Learning (AML) studio is utilized for data pre-processing, and for subsequent training, evaluation, and deployment of a machine learning model as a web service for future predictions using Excel Online. The dataset which is used for this project is “Flight Delays Data”, a built-in dataset in AML studio. The built-in “Flight Delays Data” in AML studio consists of 2,719,418 rows (observations) and 14 columns (variables). A more detailed overview for this dataset is provided in Activity 2.  
  
For this project, a machine learning model which is aimed to predict how many minutes a flight will be delayed or early is built (hence the dependent variable is ArrDelay column of the dataset). Since ArrDelay is a variable of continuous value, hence this is a regression problem and the machine learning model shall be based on regression algorithm. However, prior to training the model, some data pre-processing (eg. joining of 2 datasets, data cleansing, standardization for selected numerical variables etc.) and some data visualization (using the Execute R Script module in AML studio) shall be done. After completing the data-preprocessing and data visualization, a regression model can be built in AML studio. Finally, the trained model is published as a web service and can be shared with others, for which future predictions for ArrDelay based on other set of data can be readily made using Excel Online.

Project Technical Environment: (Describe the Microsoft Azure Machine Learning Platform)  
For this project, the following technical tools are used:  
1. Microsoft Azure Machine Learning Studio  
2. Microsoft Excel Online  
3. Microsoft Word  
  
Microsoft Azure Machine Learning (AML) studio is a workspace where we can create, build, train machine learning models. It is a drag and drop tool where we can drag the datasets and further process the analysis on that data. It offers both no-code and low-code options for projects. Codes written in Python, R or SQL for operations not offered by built-in AML modules are possible to be integrated into AML studio using the respective modules. The trained model in AML studio can be published as a web service and predictions can be made using Microsoft Excel Online. Finally, Microsoft Word is used for report preparation of this project.

1. Analytical Technique & Tools used: Describe the Analytical Technique (Regression) and Tools used in the Project  
     
   In this project, relevant modules / tools provided in AML studio are used for the following tasks:  
   a. For data-preprocessing:  
   As the original “Flight Delays Data” contains a lot of missing values and duplicate rows, the dataset shall be pre-processed in order to rectify them. This can be done by using the “Remove Duplicate Rows” and “Clean Missing Data” modules. In addition, as the original “Flight Delays Data” provides only the airport id (in digits) and without clearly specifies the respective airport locations and names, another built-in dataset “Airport Codes Dataset”, which contains the information for airport cities, states and names, will also be loaded and joined with the “Flight Delays Data” in order to produce a clearer dataset later on. Besides, before training the machine learning model, the selected numerical variables should be standardized using “Normalize Data” module (transformation method: ZScore) in order to have them on a similar scale.  
     
   b. For data visualization and statistical information:  
   Codes written in R can be integrated into AML studio environment by using the “Execute R Script” module. Those visualizations of interests (histograms, boxplots, scatterplots etc.) can be produced using either R’s built-in plots or visualization package ggplot2. Statistical information of the dataset can also be retrieved using codes written in R in a similar way.  
     
   c. For building machine learning model:  
   Prior to training the machine learning model, the cleansed data (after completing part a and b as described above) shall be split into train and test data using the “Split Data” module. Subsequently, the train data shall be used for training the model using “Train Model” module. Since the dependent variable for this project is ArrDelay, a variable of continuous value, and its known labels (actual values from historical data), hence this is a regression problem of supervised machine learning.  
   For this project, “Boosted Decision Tree Regression” is chosen as the regression algorithm and is used to train the model. After the model is trained based on the train data, it shall be scored based on the test data using the “Score Model” module. Finally, the metrics of the trained model will be evaluated by using the “Evaluate Model” module.   
     
   d. For publishing the trained model as web service:  
   Once the training of the machine learning model is done, it can be published as web service via the “Set Up Web Service” at the bottom of AML studio. Once the model is successfully deployed and shared with others, it can readily be used for future predictions using Excel Online.

1. Data Science Project Team – Roles and Responsibilities Table  
   Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions.  
   Today, data science has become an essential part of many industries, given the massive amounts of data that are produced. Its popularity has grown over the years, and companies have started implementing data science techniques to grow their business and increase customer satisfaction. As such, many companies have their own dedicated data science teams, which is responsible for helping their business leaders to make sound and data-driven business decisions.   
   The data science team of a company consists of many personnel, and each role within the team will require different specialties. In order to successfully run a data-driven business, it is crucial for the company to understand the roles and responsibilities of their data science team. The tables below summarize the required skill sets and main responsibilities for some common roles in a data science team.

A picture containing timeline

Description automatically generated

A picture containing application

Description automatically generated

1. Activity 1: Activity Summary

This project is divided into 3 parts and details of each part are elaborated in the table below:

|  |  |
| --- | --- |
| Part 1: Data Cleansing | |
| Activity 1 | Please refer page 6 for roles and responsibilities of a data science team. Please refer the details for each activity of this project provided in this section. |
| Activity 2 | “Flight Delays Data” in AML studio which is used for this project consists of 2,719,418 rows (observations) and 14 columns (variables). |
| Activity 3 | A new experiment named “DSE0322A-Lee Jack Shiang-Project” is created in AML studio. “Flight Delays Data” is loaded, and it consists of 2,719,418 rows (observations) and 14 columns (variables). The mean value of ArrDelay column is 6.6377. |
| Activity 4 | Another dataset “Airport Codes Dataset”, is loaded and joined with “Flight Delays Data” in order to show more details of airports (cities, states and names). Based on the joined dataset, it is found that “Hartsfield-Jackson Atlanta International” is the most frequently occurring destination airport in the dataset. |
| Activity 5 | Rows with matching values for all the following fields are considered duplicated and removed:  Year, Month, DayofMonth, Carrier, OriginAirportID, DestAirport ID, CRSDepTime and CRSArrTime After removing the duplicate rows, those missing values in the DepDelay and ArrDelay columns are replaced with value 0 (zero). The resultant dataset is left with 2,719,397 rows and 22 columns. The mean value of the ArrDelay column now becomes 6.5669. |
| Part 2: Data Exploration | |
| Activity 6 | For this part (Activity 6, 7, 8, and 9), codes written in R are integrated with the AML studio environment. Some functions of R, such as summary() and sd(), are utilized to obtain the statistical summary of the ArrDelay column: Minimum: -94 Mean: 6.567 Standard Deviation: 38.44812 Maximum: 1845 |
| Activity 7 | Boxplot and histogram of ArrDelay are created for visualizing its distribution. The purposes of these plots are to visualize the type of distribution (normal or skewed) of ArrDelay and if any outliers in the dataset. |
| Activity 8 | Histograms for the following columns conditioned by the ArrDel15 column are created: DepDelay, CRSArrTime, CRSDepTime, DayofMonth, DayOfWeek and Month The purpose of these facetted histograms is to visualize the distribution of the variables above, based on whether the flight was delayed by 15minutes or more. |
| Activity 9 | Scatter plots for the following columns conditioned by the ArrDel15 column are created: DepDelay, CRSArrTime, CRSDepTime, DayofMonth, DayOfWeek and Month The purpose of these conditioned scatter plots is to visualize the relationships between ArrDelay and the variables above, based on whether the flight was delayed by 15minutes or more. |
| Part 3: Machine Learning | |
| Activity 10 | A machine learning model is created, which is aimed to predict the value of ArrDelay (dependent variable) based on the following selected features (independent variables): - Month, DayofMonth, DayOfWeek, Carrier, OriginAirportID, DestAirportID, CRSDepTime, DepDelay and CRSArrTime Prior to training the model, the OriginAirportID, DestAirportID and Carrier columns are converted to Categorical. In addition, CRSDepTime, CRSArrTime and DepDelay columns are standardized using ZScore transformation method in order to make them on a similar scale. The final dataset is then split into train (70%) and test (30%) sets with random seed value of 0. The model is then trained with the train data, and the algorithm chosen for this regression problem is “Boosted Decision Tree Regression”. The trained model is then scored and evaluated based on the train and test data. The purpose of using both train and test data for the evaluation is to ensure there is no underfitting / overfitting and the trained model can perform well on both datasets. |
| Activity 11 | After finish training the model, the metrics of the trained model based on train (left) and test (right) data can be found:  It is found that the Root Mean Squared Error (RMSE) and Coefficient of Determination of the trained model based on train and test data are very close to each other. |
| Activity 12 | The trained model is ready for publish as a web service (Predictive Experiment). For this web service, only the “Scored Labels” (the predicted values of ArrDelay) column will be the output. After saving and running the Predictive Experiment, the trained model can be deployed and shared with others for future predictions of ArrDelay based on other dataset using Excel Online. |

1. Activity 2: Flight DataSet- An Overview

The dataset which is used for this project is the “Flight Delays Data”, a built-in dataset in AML studio. It consists of 2,719,418 rows (observations) and 14 columns (variables) of air flight data collected in 2013 from various airports in the United States.

Chart

Description automatically generated

The 14 columns of the dataset are:  
Year: Year  
Month: Month  
DayofMonth: Day of the month  
DayOfWeek: Day of the week  
Carrier: Code assigned by IATA (International Air Transport Association) and commonly used to identify a carrier  
OriginAirportID: An identification number assigned by US DOT (Department of Transportation) to identify a unique airport (the flight’s origin)  
DestAirportID: An identification number assigned by US DOT to identify a unique airport (the flight’s destination)  
CRSDepTime: The CRS departure time in local time (hhmm)  
DepDelay: Difference in minutes between the scheduled and actual departure times. Early departures show negative numbers  
DepDel15: A Boolean value indicating whether the departure was delayed by 15 minutes or more (1 = Departure was delayed)  
CRSArrTime: CRS arrival time in local time (hhmm)  
ArrDelay: Difference in minutes between the scheduled and actual arrival times. Early arrivals show negative numbers  
ArrDel15: A Boolean value indicating whether the arrival was delayed by 15 minutes or more (1 = Arrival was delayed)  
Cancelled: A Boolean value indicating whether the arrival flight was cancelled (1 = Flight was cancelled)  
  
As the aim of the machine learning model of this project is to predict how many minutes a flight will be delayed or early, the dependent variable is the ArrDelay column of the dataset.

1. Screen-shots of each task of Activity 3: Create New Experiment

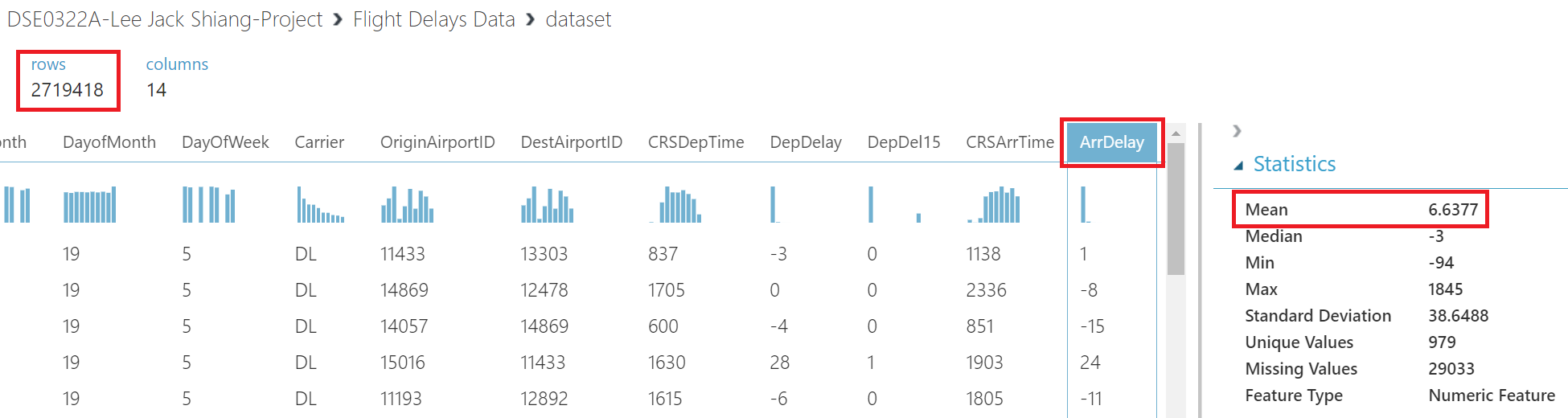
Task 1: Sign in to Azure Machine Learning Workspace

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.

Graphical user interface, text, application, chat or text message

Description automatically generated



Task 4: Answer the following questions:

- How many rows are in the dataset?

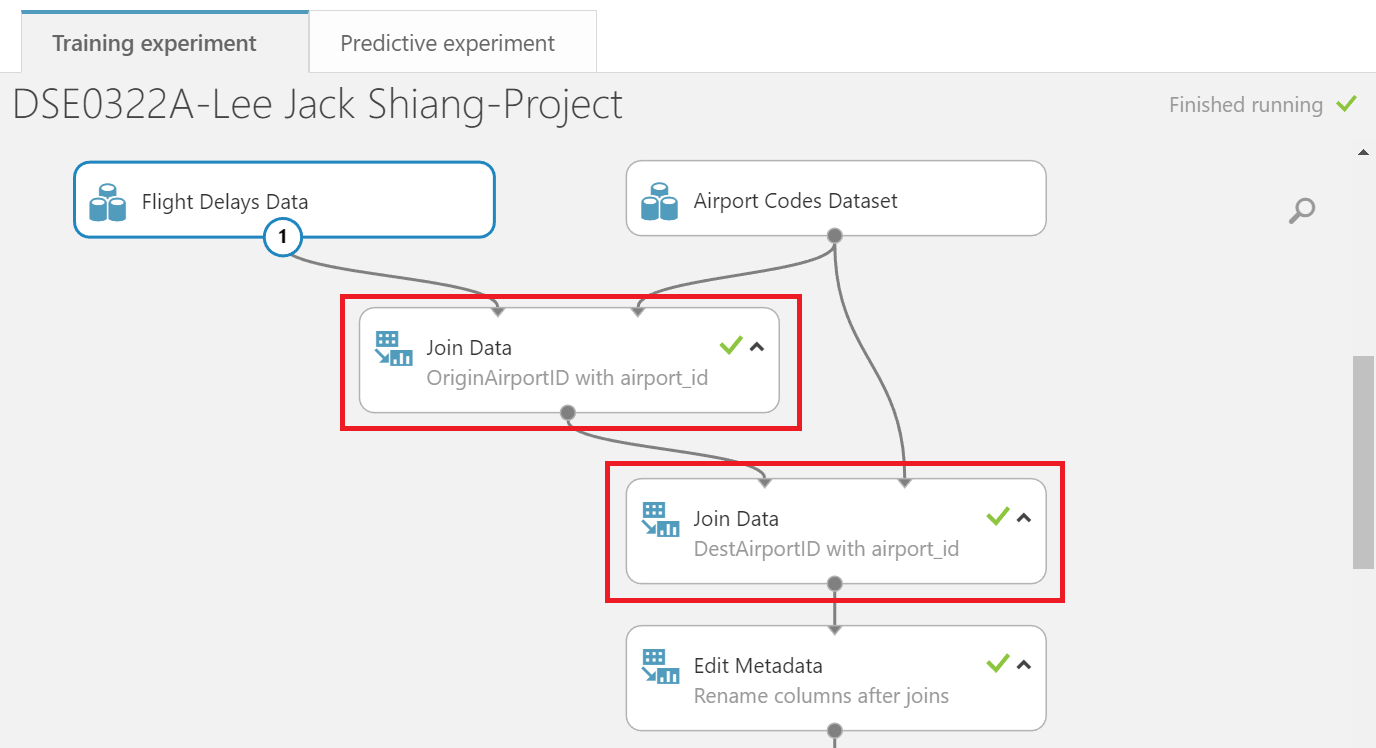
2,719,418

- What is the mean value of the ArrDelay column?

6.6377

1. Screen-shots of each task of Activity 4: Join the Airport Codes Dataset

Task 1: Create two joins; one to join the origin airport and one to join the destination airport. Use built-in Azure Machine Learning Join Data modules to accomplish this, or use custom SQL or R code. The resulting joined dataset should contain the original flight data as well as the name, city, and state for both the origin and destination airports.



Task 2: After joining the datasets, visualize the joined data.



Task 3: Answer the following questions: Which is the most frequently occurring destination airport in the dataset.

Hartsfield-Jackson Atlanta International

1. Screen-shots of each task of Activity 5: Remove Duplicates and Replace Missing Values

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

- CRS Arr Time

Use the built-in Azure Machine Learning module, or custom SQL or R script to accomplish this.

A picture containing graphical user interface

Description automatically generated Graphical user interface, text, application, email

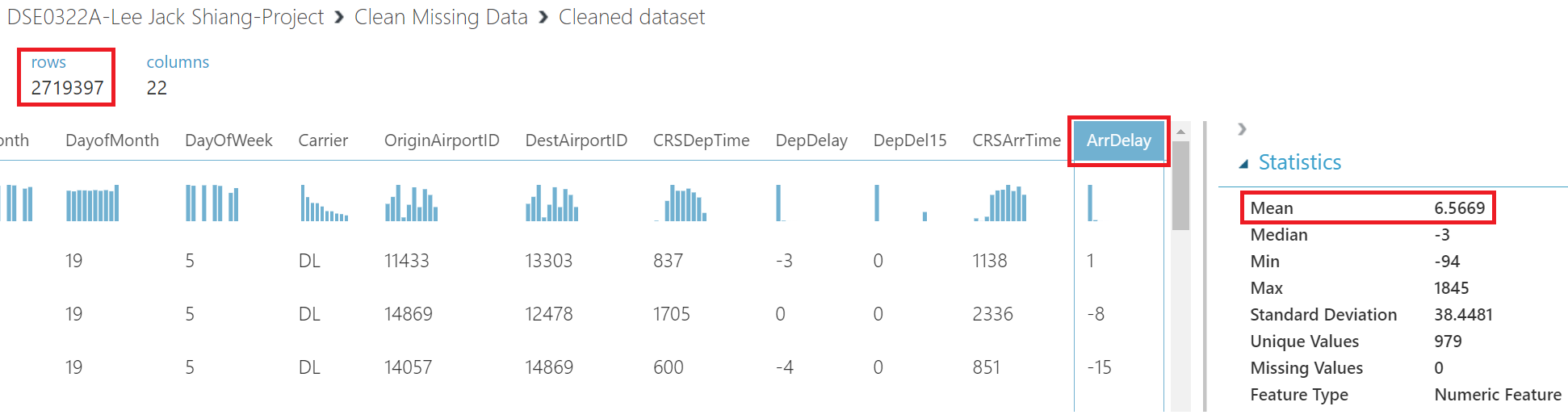
Description automatically generated

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 Remove Duplicate Rows module, or custom SQL or R script to accomplish this.

Graphical user interface, text, application

Description automatically generated Table

Description automatically generated



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

- How many rows remain in the dataset?

2,719,397

- What is the mean value of the ArrDelay column?

6.5669

It is required to handle those duplicate rows and missing values of a dataset prior to training the machine learning model as:

1. Many machine learning algorithms do not work with missing values. To avoid any errors or unexpected outcomes during model training, those missing values must be handled beforehand.
2. Duplicate rows mean feeding the same information to the model again and again when it is being trained. This will result in a biased and overfitted model.

1. Screen-shots of each task of Activity 6: View Summary Statistics

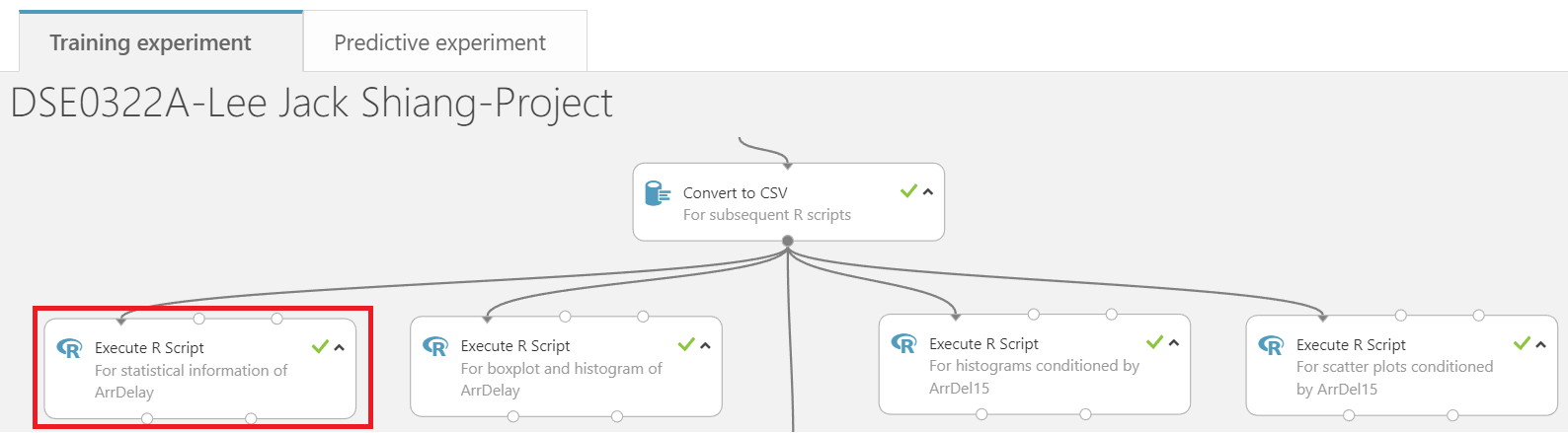
Task 1: Open the Azure Machine Learning experiment that was used to cleanse the flight data.

Task 2: Add a Convert to CSV module to the last existing module and run the experiment

Graphical user interface, text, application, chat or text message

Description automatically generated

Task 3: Right-click the output of the Convert to CSV module and open the dataset in a new notebook using R



Task 4: After the notebook has opened, run the existing cells (which are generated automatically) to load and view the DataFrame. Tip: Rename your notebook to something more meaningful (like "Explore Flight Data"). This will help you find it again if you want to reopen it later

Graphical user interface, text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated Table

Description automatically generated

Task 5: Use the R summary and sd functions to display summary statistics for all columns in the DataFrame containing the flights data.

A picture containing text

Description automatically generated

Task 6: Determine the following summary values for the ArrDelay column:

- Minimum

-94.0

- Mean

6.567

- Standard Deviation

38.44812

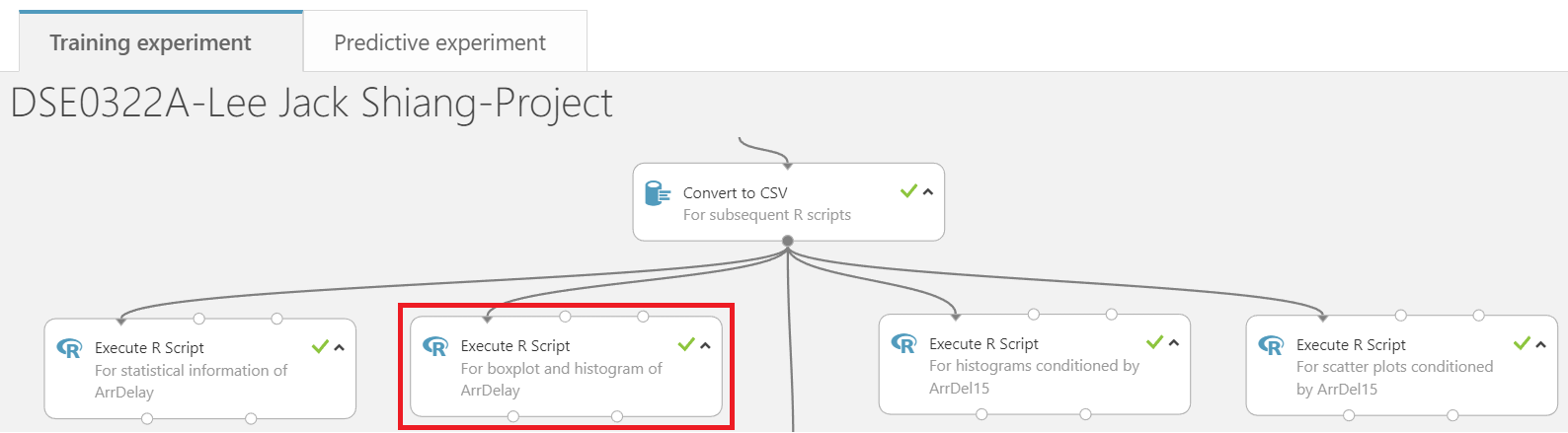
- Maximum

1845.0

1. Screen-shots of each task of Activity 7: View ArrDelay Distribution

Task 1: Determine the range and distribution of values for this value.

Task 2: Explore the range and distribution of values in the ArrDelay column after creating a plot that shows a box plot and a histogram of this value. The histogram should display the values in 30 bins.



R Script:

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, diagram

Description automatically generated with medium confidence

Task 3: Answer the following question and enter the answer:

Based on the data visualization, which three of the following statements accurately reflect the distribution of ArrDelay values?

- The median, first quartile, and third quartile are all fairly close to 0, indicating that most flights arrive close to their scheduled time. ✓

~~- The distribution is evenly spread across the range of values, indicating that flights are equally likely to be early, on-time, or late.~~ ❌

~~- The median indicates that the majority of flights are 500 minutes late or more.~~ ❌

- The range of arrival times ranges extensively, with some flights arriving as much as 1500 minutes late. ✓

- The distribution is right-skewed, so there is a higher range of values for late flights than for early flights. ✓

~~- The distribution is left-skewed, so there is a higher range of values for early flights than for late flights~~❌

1. Screen-shots of each task of Activity 8: Use Histograms to Compare Numeric Columns

Task 1: To explore how these values might be related to arrival delay, plot histograms conditioned by the ArrDel15column, which is a binary column indicating whether a flight arrived 15 or more minutes late.

Write code to generate conditioned histograms for the following columns, conditioned by the ArrDel15 column:

- DepDelay

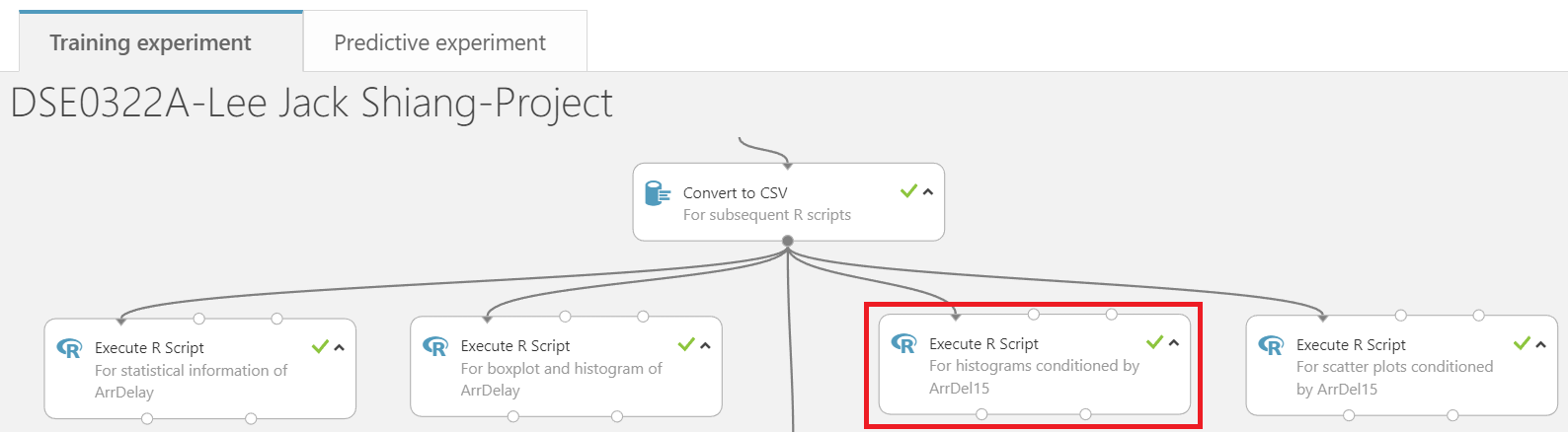
- CRSArrTime

- CRSDepTime

- DayofMonth

- DayOfWeek

- Month



R Script:

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Task 2: Answer the following question:

Based on the conditioned histograms, which three of the following statements are true? - There are significantly more flights that are less than 15 minutes late than there are flights that are 15 minutes late or more. ✓

~~- Late flights tend to occur more frequently at the beginning of the month.~~ ❌

- Flights that are 15 minutes or more late tend to have a higher DepDelay value than flights that are on-time. ✓

- Late flights tend to occur more frequently for flights with a CRSArrTime that is later in the day, the highest volume of delayed flights scheduled to arrive between 3pm (1500 hours) and 8pm (2000 hours) ✓

~~- The relative distribution of late flights varies significantly from that of on-time flights based on the day of the week.~~ ❌

1. Screen-shots of each task of Activity 9: Use Scatter Plots to Compare Numeric Columns

Task 1: Write code to generate conditioned scatter plots for the following columns, conditioned by the ArrDel15 column using different colors for values of 0 and 1:

- DepDelay

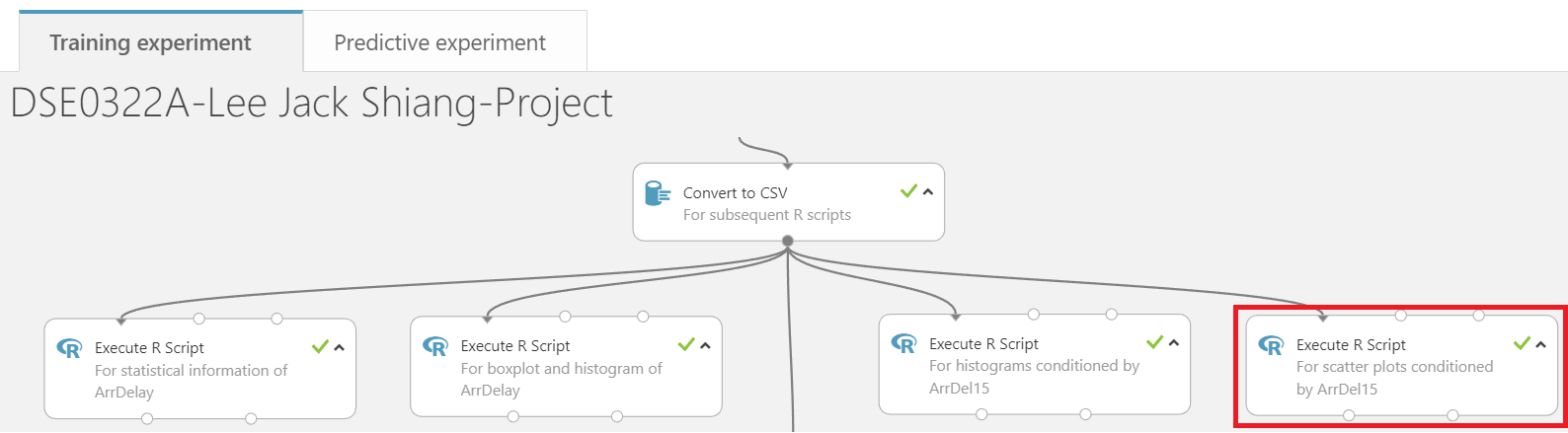
- CRSArrTime

- CRSDepTime

- DayofMonth

- DayOfWeek

- Month



R Script:

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Chart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Task 2: Answer the following question and enter the answer on EdX:

Based on the conditioned scatter plots, which two of the following statements are true? ~~- There is a clear relationship between ArrDelay and month. Later months of the year show markedly longer delays.~~ ❌

- There is a near-linear relatonship between DepDelay and ArrDelay for late flights. As departure delay increases, so does arrival delay. ✓

~~- There is a clear correlation between ArrDelay and DayofMonth. Earlier days of the month show markedly longer delays.~~ ❌

- There is an apparent relationship between ArrDelay and CRSDepTime. Flights that depart early in the morning are typically less delayed than flights that are scheduled to depart after around 5am (0500 hours), at which time delays tend to get significantly longer. Delays then gradually get shorter as the day progresses ✓

1. Screen-shots of each task of Activity 10: Train a Regression Model

- Task 1: Return to the Azure Machine Learning experiment you created in Part 1.

- 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.

Graphical user interface, application, Teams

Description automatically generated

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

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

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

Graphical user interface, application

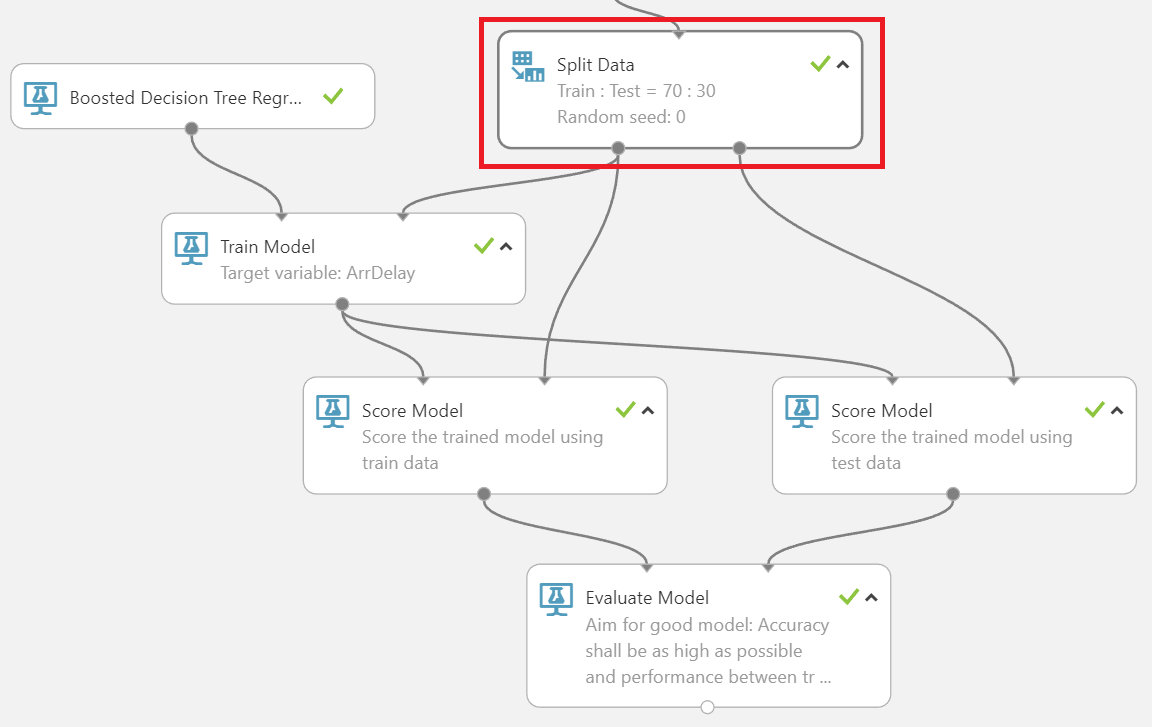
Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

The reason to standardize these numerical variables is to have them on a similar scale. Variables that are measured at different scales do not contribute equally when training the machine learning model and this may end up with a biased model (eg. a variable that ranges between 0 and 1000 will outweigh a variable that ranges between 0 and 1).

- 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.

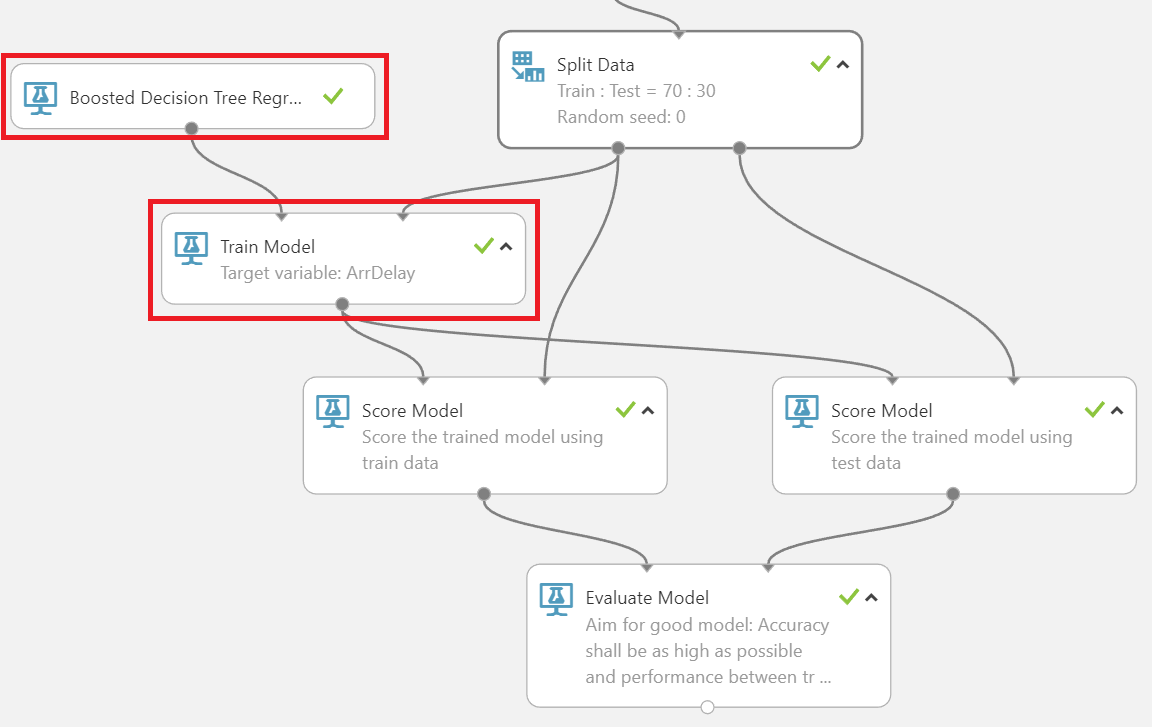


Graphical user interface, text, application

Description automatically generated

The purpose of the train-test split is to estimate the performance of a machine learning model when it is used for making predictions based on real world data for which the trained model has not seen before. In addition, train-test split can also help to minimize overfitting and the accuracy of the model based on test data can be directly compared with the accuracy based on train data. If the model’s accuracy on the test data is much worse than that of on the train data, then the model is overfitted during training. This check for overfitting is not possible if there is no train-test split.

- 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.



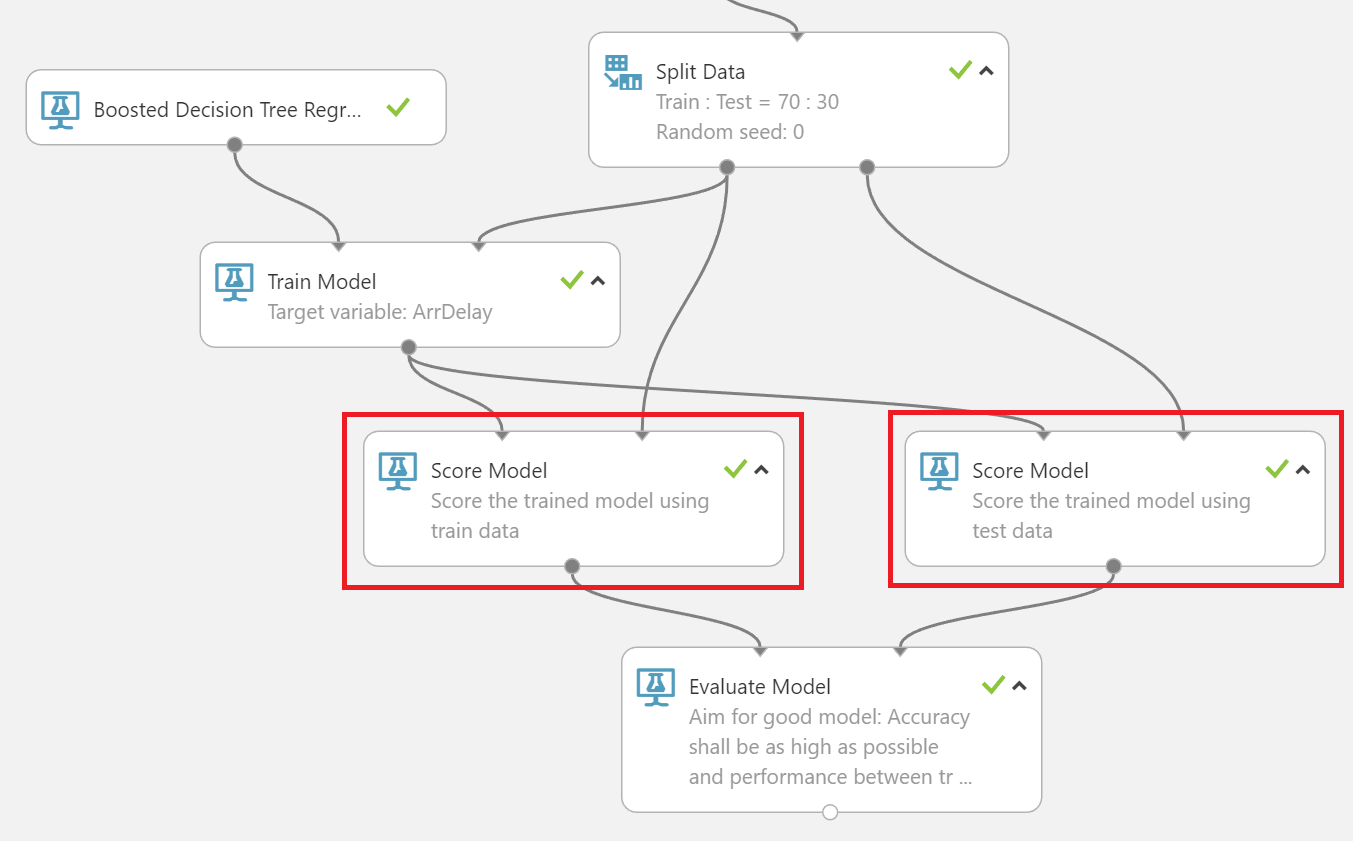
Graphical user interface, text, application, email

Description automatically generated

Text, application

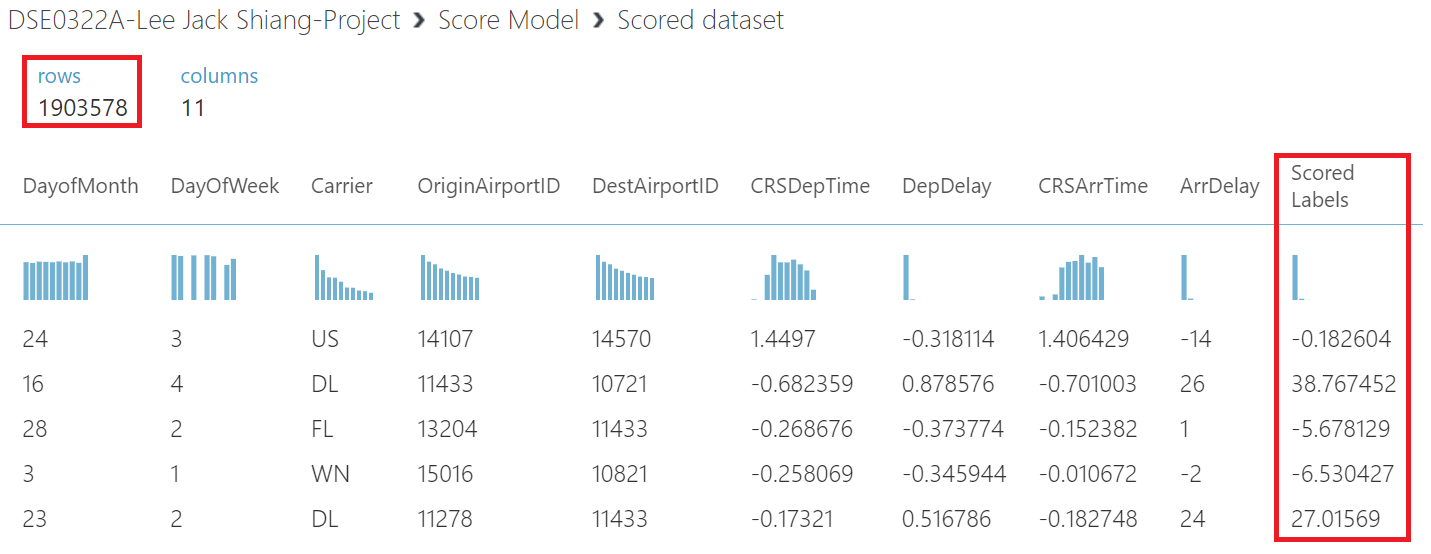
Description automatically generated

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

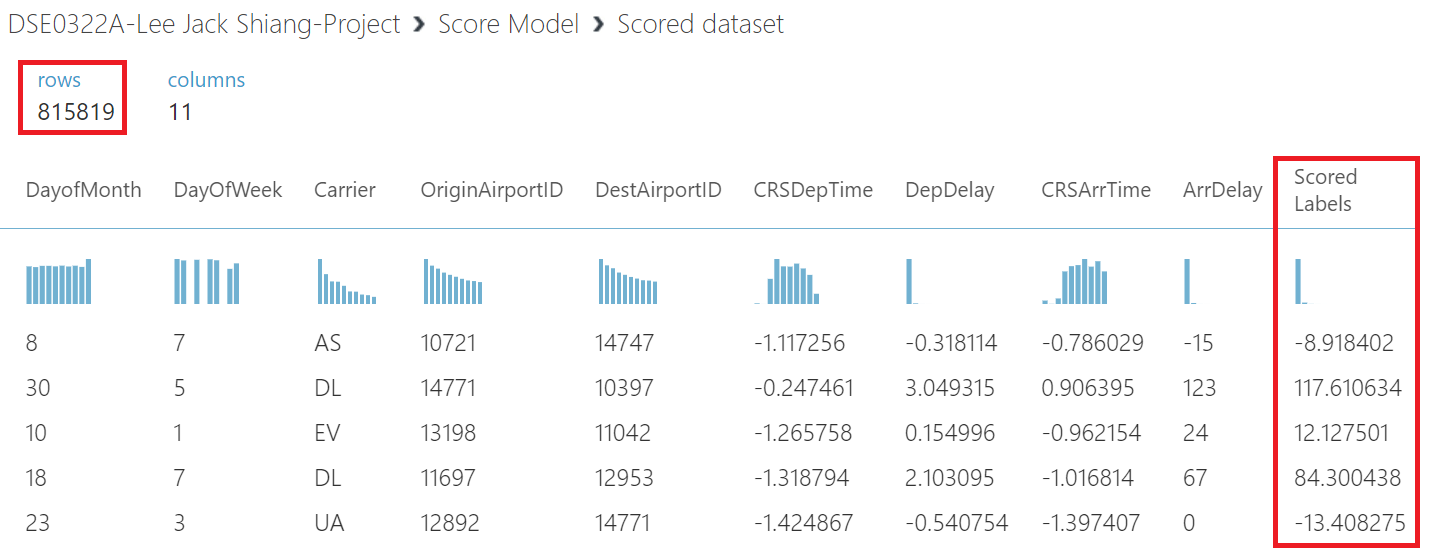


Note: Score Model module is added for both train and test data to ensure there is no underfitting / overfitting and the trained model can perform well on both datasets.

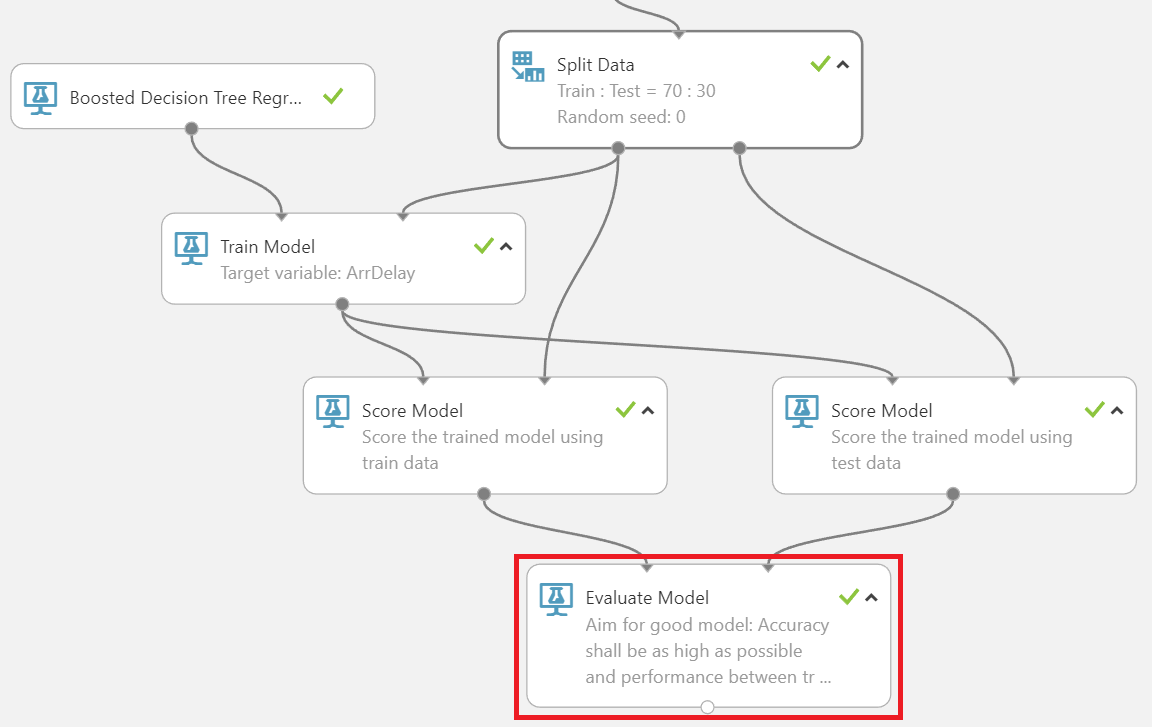
Score Model based on train data:



Score Model based on test data:



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

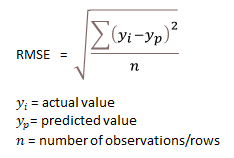


1. Screen-shots of each task of Activity 11: Test and Evaluate the Mode

Task 1: Run the experiment,

Task 2: When it has finished, visualize the output of the Evaluate Model module. There are numerous metrics that you can use to evaluate a regression model, including the Root Mean Squared Error (RMSE), which indicates the mean variance between predicted and actual label values, in this case, the number of minutes on average by which predicted flight delays vary from actual flight delays.

RMSE is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance and it can be expressed as



The coefficient of determination (R² or r-squared) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable.

Text

Description automatically generated

Text, application

Description automatically generated Text

Description automatically generated  
  
  
  
  
If the problem is multivariate (more than 1 independent variable), adjusted R² shall be used and it is given by the expression below:  
  
Text

Description automatically generated with medium confidence

Metrics for model evaluated based on train (left) and test (right) data:

A picture containing graphical user interface

Description automatically generated

Note: For a good model, its accuracy shall be as high as possible and performance between train and test data shall be as close as possible.

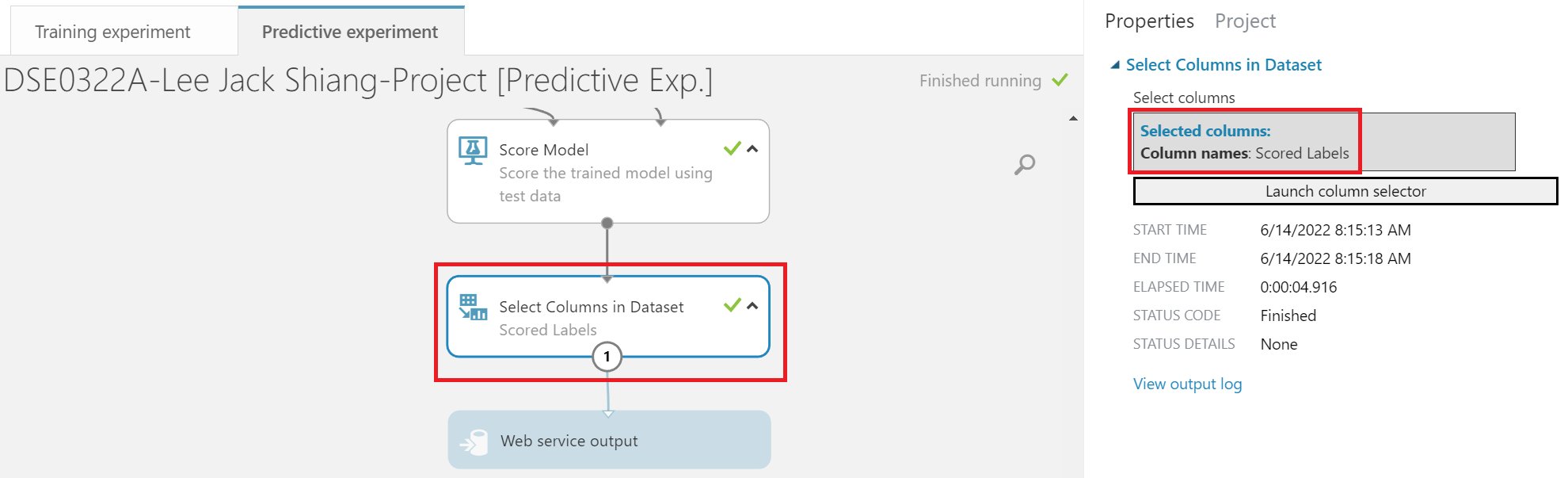
Task 3: Determine the Root Mean Squared Error for the model.

Root Mean Squared Error for the model (based on test data): 12.778422

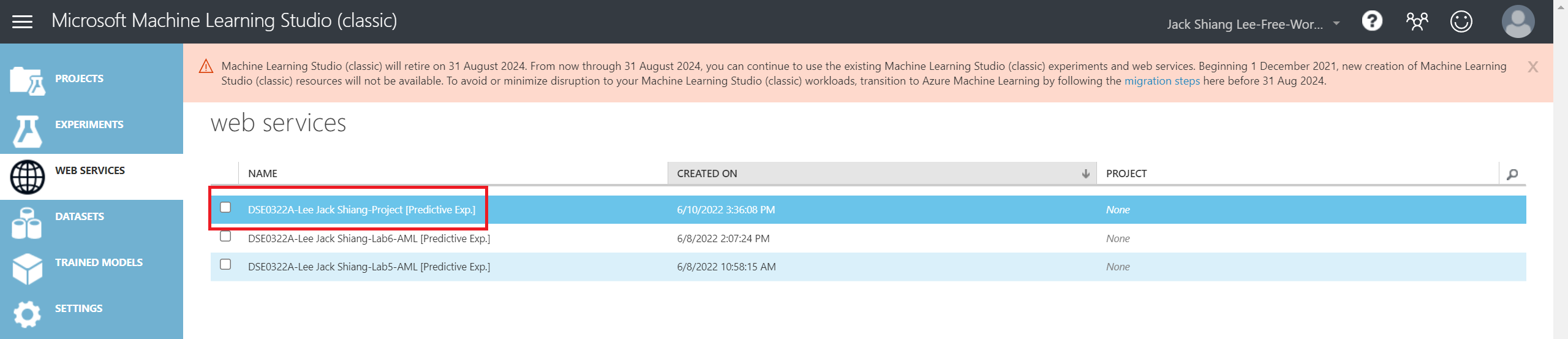
1. Screen-shots of each task of Activity 12: Publish and Use the Model

Task 1: Set up the experiment as a web service, creating a predictive experiment (if the option to do this is not available, save and re-run the experiment).

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.



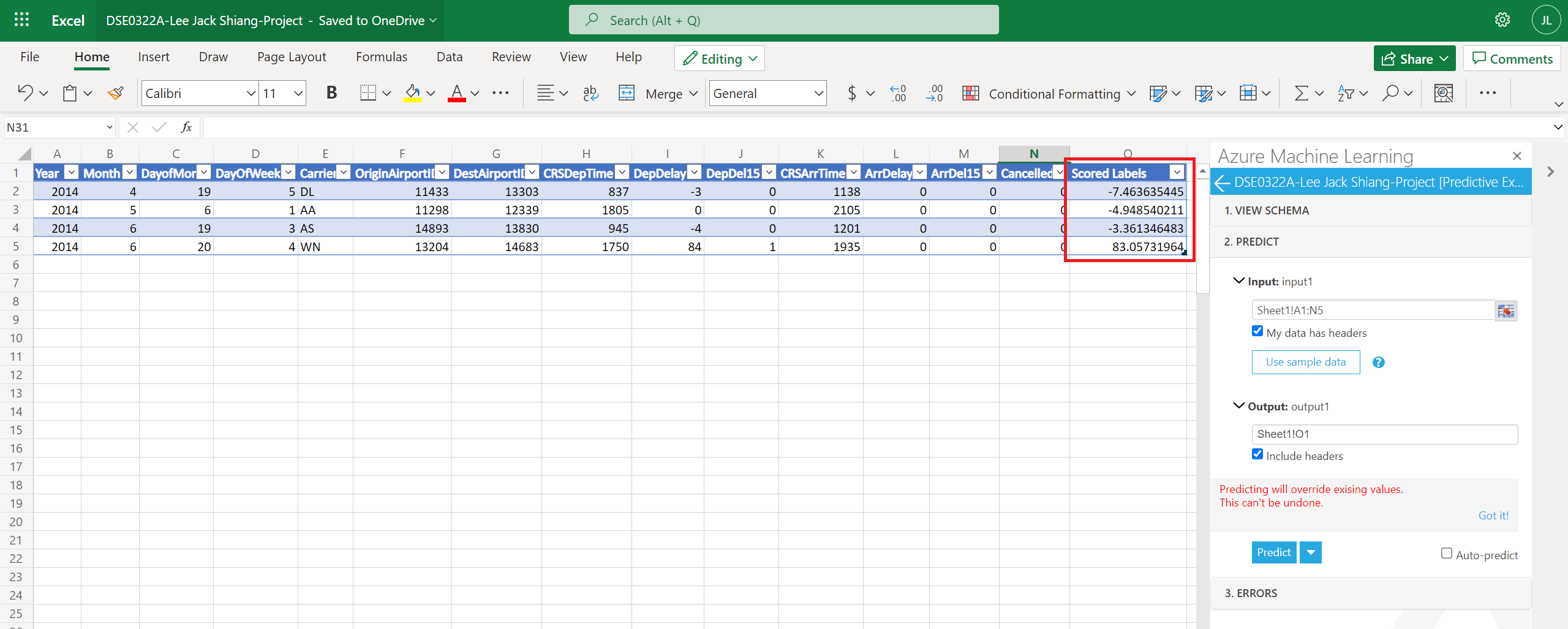
Graphical user interface, text, application, Word

Description automatically generated

Task 4: Use Excel to test the web service, and generate predicted values for the input rows below

Graphical user interface, application

Description automatically generated



Task 5: Enter the values predicted by your model for each input row:

- Input Row 1:

-7.46363544464111

- Input Row 2:

-4.94854021072388

- Input Row 3:

-3.36134648323059

- Input Row 4:

83.0573196411133

1. Conclusion

In this project, it has demonstrated a typical workflow of building a machine learning model using AML studio. From dataset import, data-preprocessing, data visualization, training of the machine learning model, to final model deployment into production, it is possible to build a machine learning model just by using AML studio’s drag and drop tools without any coding (data visualization part can be done using other software, such as PowerBI, Tableau, Excel etc). However, it should be noted that not all possible operations are available in AML studio’s modules. Hence, to truly leverage the power of machine learning, one should also possess certain level of coding skills (eg. in Python and R) in order to have full control when building the machine learning model.

1. Annexure – 01

This Annexure shows the screenshots of the full workflow for building the machine learning model in AML studio (Training experiment).

Graphical user interface, application

Description automatically generated

Graphical user interface, diagram

Description automatically generated

1. Annexure – 02

This Annexure shows the screenshots of the full workflow for the deployed machine learning model into web service in AML studio (Predictive experiment).

Graphical user interface

Description automatically generated

Graphical user interface, application

Description automatically generated

REQUEST/RESPONSE link and API key of the model for Azure Machine Learning Add-ins in Excel Online:

REQUEST/RESPONSE link:

<https://studio.azureml.net/apihelp/workspaces/6587b92216804c8dbe342b49760300e2/webservices/8ea2e37b6e3d4376a58b1e810f26f508/endpoints/6342e078c48f449d845961d22d6175e8/score>

API key:

PpHtWgBxCevCDe9T+ubwyRrPS40ykYv6ZstFgVOd174VlMoSmCBSLeJfjnHRm5tkrK6LEwlEBCOq6nvek1YeRQ==