Project Implementation

I have taken the idea of this project from azure gallery wherein I use predictive modelling to find out whether the arrival of a scheduled passenger flight will be delayed based on weather data.

As we know, Classification is generally a supervised learning problem. This project is a binary classification task predicting two classes -- whether the flight will be delayed, or whether it will be on time.

I built an experiment using Azure ML Studio in which I trained a model using 1 year of historic flight data.

Below are the basic steps that I have implemented in an experiment in Azure ML Studio-

- Importing data
- Data pre processing
- Join Datasets
- Train the model using algorithm
- Score & Evaluate model

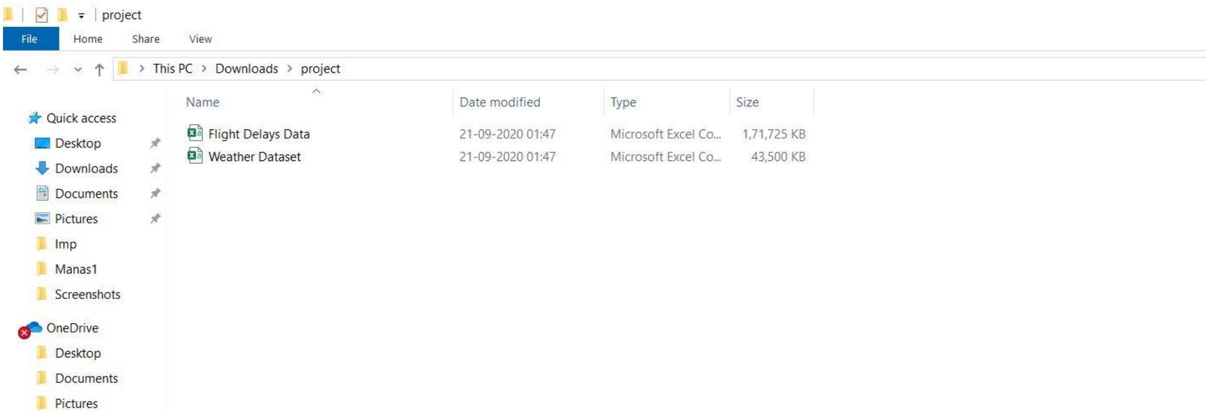
Importing data

I have downloaded 2 datasets namely flights & weather from 'U.S. Department of Transportation'. The dataset contains flight delay data since Jan'2019. Cancelled flights was relabelled as delayed by more than 15 mins. These datasets were saved as .csv in my local machine.

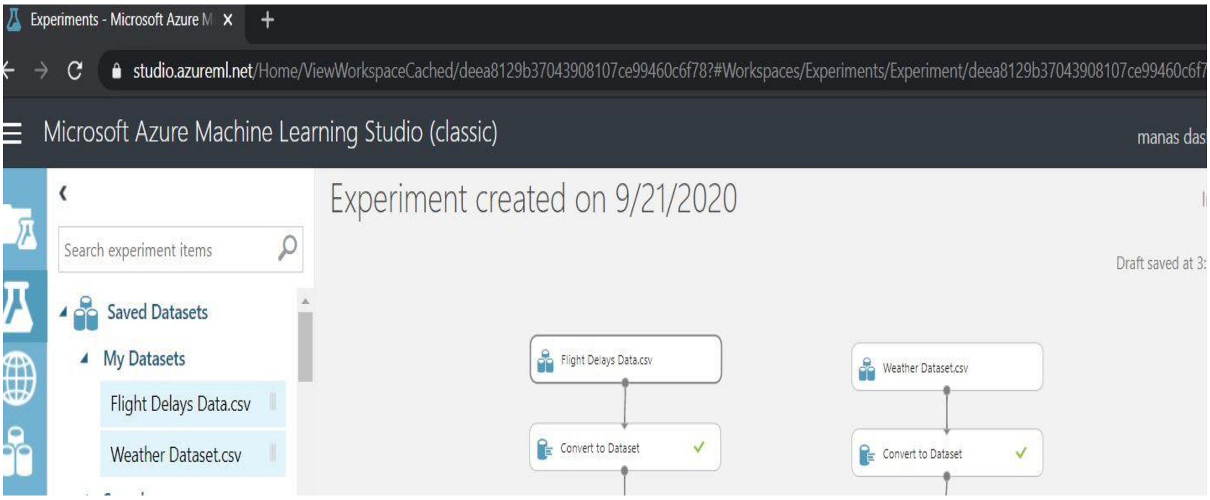
Columns below-

Flights.csv: Year, Month, DayofMonth, DayOfWeek, Carrier, OriginAirportID, DestAirportID, CRSDepTime, DepDelay, DepDel15, CRSArrTime, ArrDelay, ArrDel15, and Cancelled.

Weather.csv: AirportID, Year, Month, Day, Time, TimeZone, SkyCondition, Visibility, WeatherType, DryBulbFarenheit, DryBulbCelsius, WetBulbFarenheit, WetBulbCelsius, DewPointFarenheit, DewPointCelsius, RelativeHumidity, WindSpeed, WindDirection, ValueForWindCharacter, StationPressure, PressureTendency, PressureChange, SeaLevelPressure, RecordType, HourlyPrecip, Altimeter

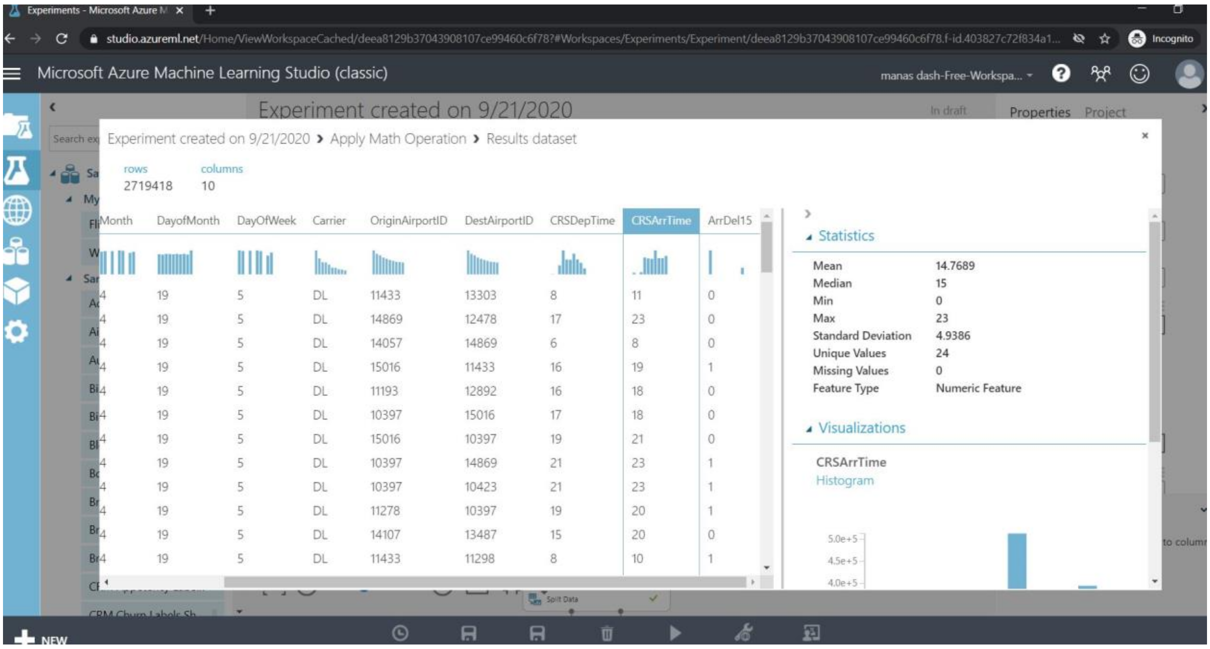


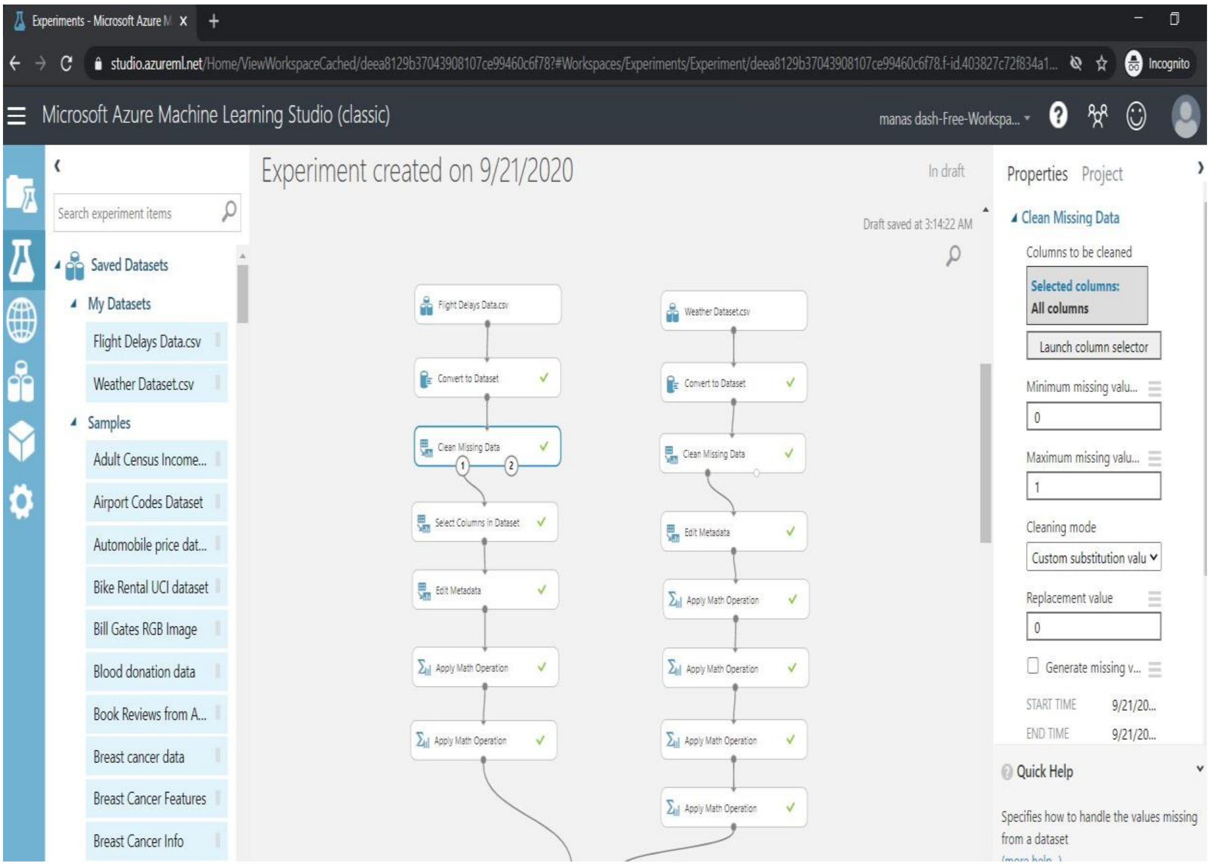
I have used Azure ML studio to convert these csv files to datasets after creating an experiment.



Data pre processing

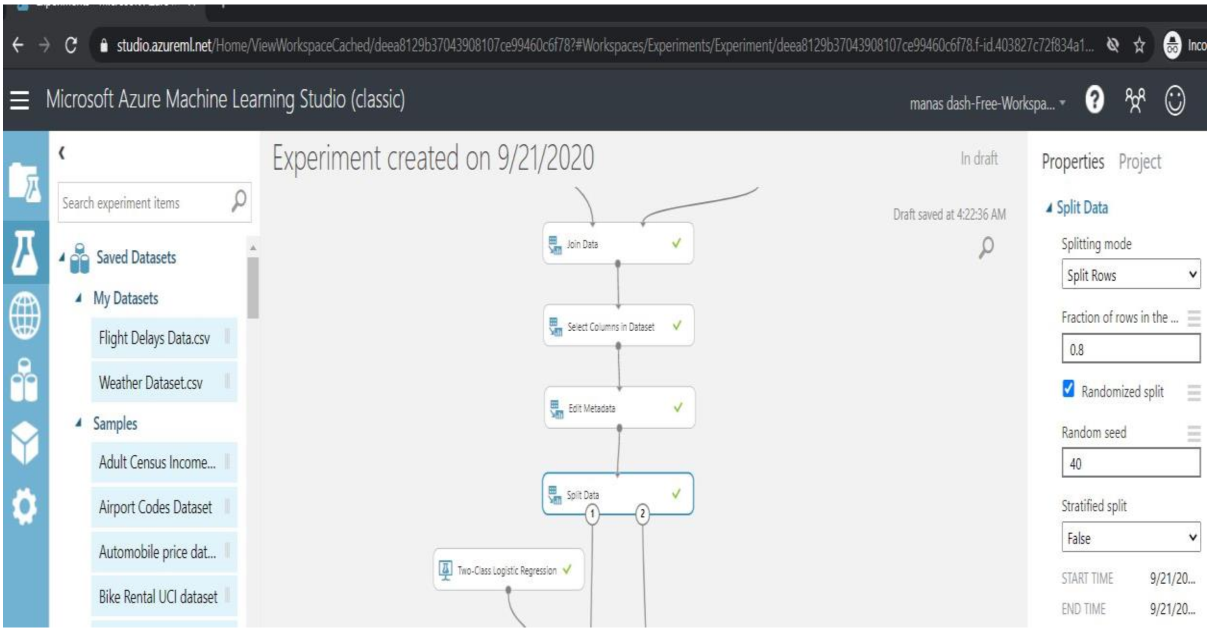
After scrubbing missing values from both the datasets, I excluded 4 columns Cancelled, ArrDelay, DepDelay, DepDel15 to eliminate target leaks. 'Edit metadata' feature was used to convert numerical columns such as OriginAirportID, DestAirportID, Carrier to categorical. A bunch of 'Math operations' were applied to extract/transform the respective time fields. Below are the screenshots with absolute values and the experiment so far.





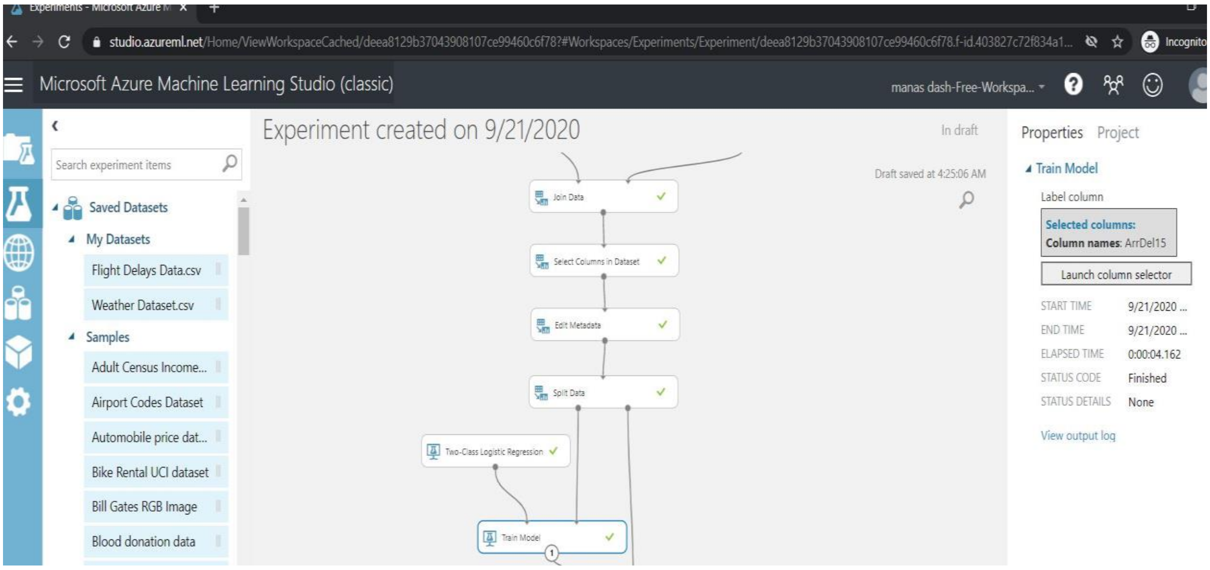
Join Datasets

I used inner join to join both the datasets. I did an 80:20 randomized split of data and used the two class logistic regression algorithm to train the model.



Train the model

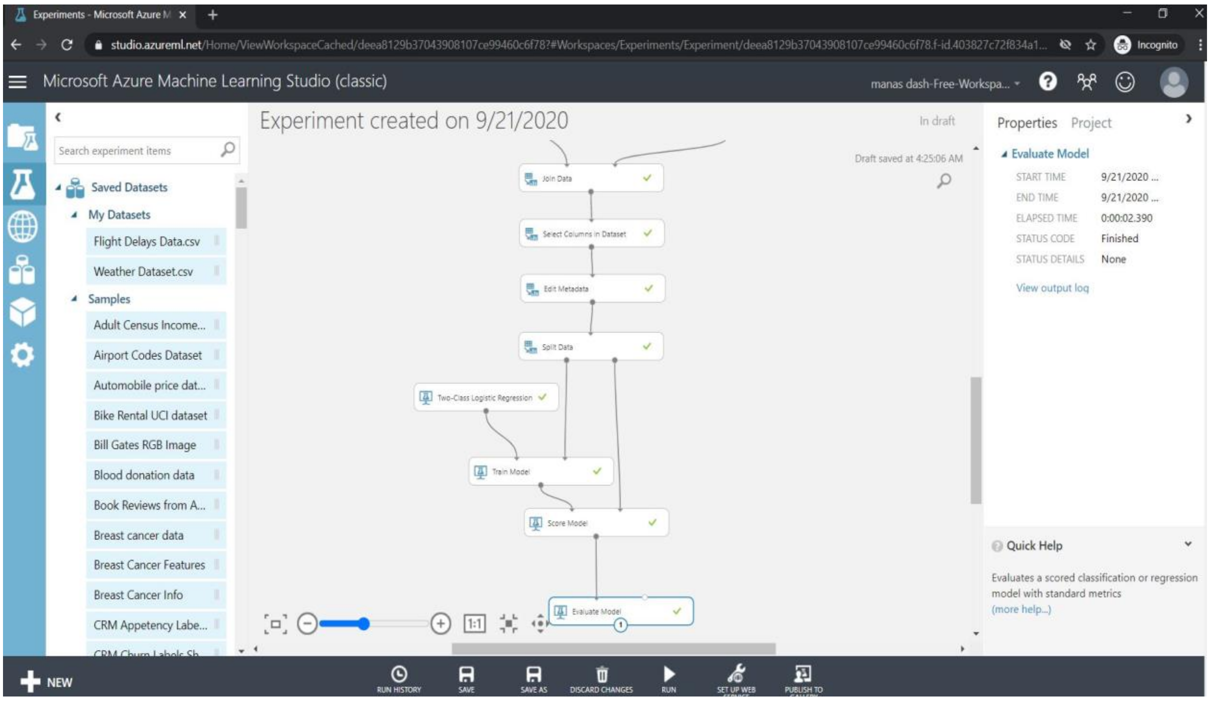
The model was trained using logistic regression on the column of arrival delay.



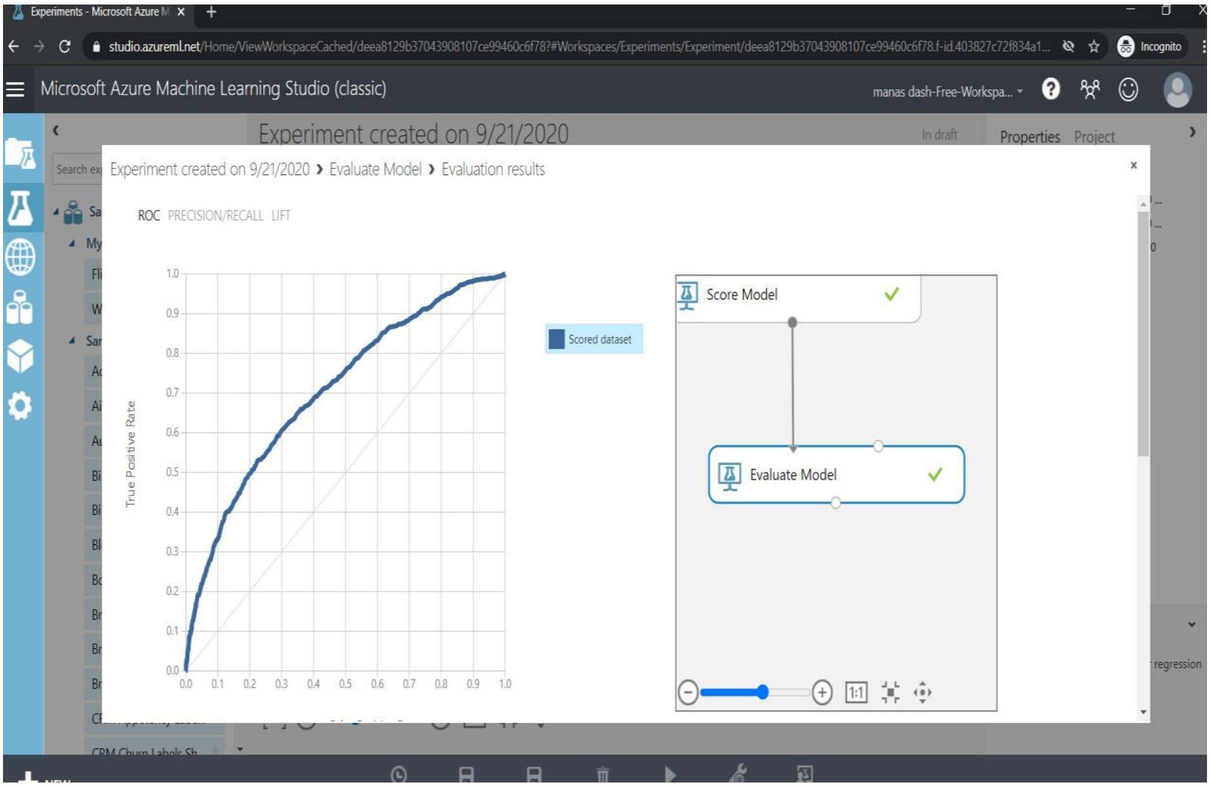
Score & Evaluate model

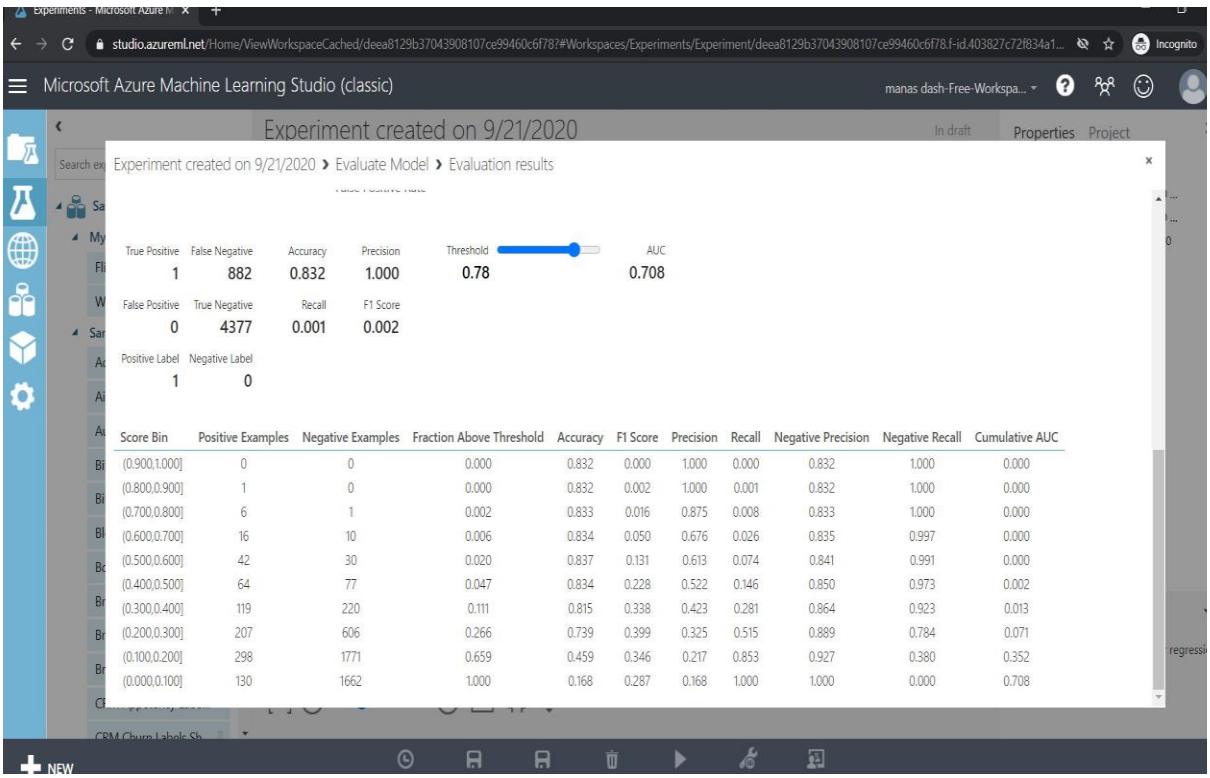
The various features used were the weather conditions at the arrival and destination airports, departure and arrival times, the airline carrier, the day of month, and the day of the week.

The result of the experiment is a trained classification model that can be used to score new samples to make predictions. Scores generated from the trained models was used to Evaluate Model module to analyse and compare the quality of the models.



The model resulted in AUC of 0.708.





Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	0	0	0.000	0.832	0.000	1.000	0.000	0.832	1.000	0.000
(0.800,0.900]	1	0	0.000	0.832	0.002	1.000	0.001	0.832	1.000	0.000
(0.700,0.800]	6	1	0.002	0.833	0.016	0.875	0.008	0.833	1.000	0.000
(0.600,0.700]	16	10	0.006	0.834	0.050	0.676	0.026	0.835	0.997	0.000
(0.500,0.600]	42	30	0.020	0.837	0.131	0.613	0.074	0.841	0.991	0.000
(0.400,0.500]	64	77	0.047	0.834	0.228	0.522	0.146	0.850	0.973	0.002
(0.300,0.400]	119	220	0.111	0.815	0.338	0.423	0.281	0.864	0.923	0.013
(0.200,0.300]	207	606	0.266	0.739	0.399	0.325	0.515	0.889	0.784	0.071
(0.100,0.200]	298	1771	0.659	0.459	0.346	0.217	0.853	0.927	0.380	0.352
(0.000,0.100]	130	1662	1.000	0.168	0.287	0.168	1.000	1.000	0.000	0.708