Data Analytics Kemet's Report

Formula AI Hackathon '22
22 Feb 2022
Kemet Team

Team Members:
Daniel Joseph
Mohamed Ibrahim
Verina George
Mark Ehab
Nour Almostafa

INTRODUCTION

The following report is to document the work of the Kemet team participating in Hackmakers Formula AI Hackathon. Specifically, the Data Analytics theme.

The team was presented with historical weather data from the RedBull Racing eSports team and was required to develop an Artificial Intelligence model that is able to make accurate weather predictions / forecasts.

The dataset had undergone some data processing and cleaning procedures, that made the dataset more readable. Multiple models were created and trained using 80% of the said dataset, then testing on the remaining 20% reeled some interesting results.

By comparing the models, our ANN was the best choice.

OBJECTIVE

What we are working on:

- Creating an accurate weather prediction model for the F1 2021 videogame.
- F1 2021 videogame, the official Formula 1 videogame, is developed by Codemasters, and uses a physics engine that behaves like the real world.

The team's objective was:

- Deal with the large dataset presented to the team and have our analysis of data be as beneficial as possible.
- Clean then preprocess the analyzed data, and choose the important features according to our analysis of data.
- Begin testing with different models to compare between their behavior with this set of data.
- Choose the most suitable model.

The required output of the model:

- **Rain Percentage.** Probability that it will rain.
- **Weather Type.**
  - Clear = 0
  - light cloud = 1
  - overcast = 2
  - light rain = 3
  - heavy rain = 4
  - storm = 5

DATASET CLEANING AND PREPROCESSING

Inspecting the **correlation matrix** of each feature:

	M_WEATHER	M_RAIN_PERCENTAGE
M_GAME_MINOR_VERSION	-0.035001	-0.118432
M_SESSION_UID	-0.067724	-0.177354
M_SESSION_TIME	-0.007734	0.050886
M_FRAME_IDENTIFIER	-0.037441	-0.006011
M_PLAYER_CAR_INDEX	-0.265795	0.139027
M_BRAKING_ASSIST	0.128744	0.226468
M_SESSION_LINK_IDENTIFIER	0.09437	0.274133
M_PIT_RELEASE_ASSIST	0.202819	0.237467
TIMESTAMP	-0.118416	-0.164248
M_PIT_STOP_WINDOW_IDEAL_LAP	0.158741	-0.034702
M_TRACK_TEMPERATURE	-0.614053	0.121007
M_TRACK_LENGTH	0.389966	0.189765
M_GAME_PAUSED	0.050949	0.009488
M_FORECAST_ACCURACY	-0.062449	0.11074
M_AIR_TEMPERATURE	-0.379741	0.152132
M_NUM_WEATHER_FORECAST_SAMPLES	-0.018511	0.36706
M_TRACK_ID	-0.393741	-0.200091
M_ERSASSIST	0.202819	0.237467
M_FORMULA	-0.023969	-0.014516
M_SEASON_LINK_IDENTIFIER	0.09437	0.274133
M_PIT_ASSIST	0.202819	0.237467
M_GEARBOX_ASSIST	0.202819	0.237467
M_SESSION_TYPE	0.022748	-0.359983
M_SPECTATOR_CAR_INDEX	0.017726	0.009337
M_PIT_STOP_WINDOW_LATEST_LAP	0.160691	-0.034834
M_WEEKEND_LINK_IDENTIFIER	0.09437	0.274133
M_DYNAMIC_RACING_LINE_TYPE	0.202819	0.237467
M_SESSION_TIME_LEFT	-0.019687	0.067034
M_SESSION_DURATION	0.105859	0.062526
M_PIT_STOP_REJOIN_POSITION	0.056602	-0.089815
M_WEATHER_FORECAST_SAMPLES_M_SESSION_TYPE	0.005031	0.474839
M_TIME_OFFSET	0.005202	0.489043
M_WEATHER_FORECAST_SAMPLES_M_WEATHER	0.092816	0.8675
M_WEATHER_FORECAST_SAMPLES_M_TRACK_TEMPERATURE	-0.022552	0.575722
M_TRACK_TEMPERATURE_CHANGE	-0.016857	0.530948
M_WEATHER_FORECAST_SAMPLES_M_AIR_TEMPERATURE	-0.028636	0.570559
M_AIR_TEMPERATURE_CHANGE	-0.019445	0.535547
M_RAIN_PERCENTAGE	0.088991	1
M_WEATHER	1	0.088991
M_AI_DIFFICULTY	-0.245912	-0.06361
M_PIT_SPEED_LIMIT	-0.139448	0.13472
M_NETWORK_GAME	0.204437	0.047039
M_TOTAL_LAPS	-0.091284	0.098738
M_STEERING_ASSIST	-0.011431	0.281576
M_IS_SPECTATING	-0.017726	-0.009348
M_DYNAMIC_RACING_LINE	0.202819	0.237467
M_DRSASSIST	0.202819	0.237467
M_NUM_MARSHAL_ZONES	0.188183	0.053465

```
Data original Size : 3572328 rows x 59 columns
```

Dataset statistics

Number of variables	59
Number of observations	3572328
Missing cells	17471851
Missing cells (%)	8.3%
Total size in memory	1.6 GiB
Average record size in memory	472.0 B

The **cleaning process** consists of 6 main steps:

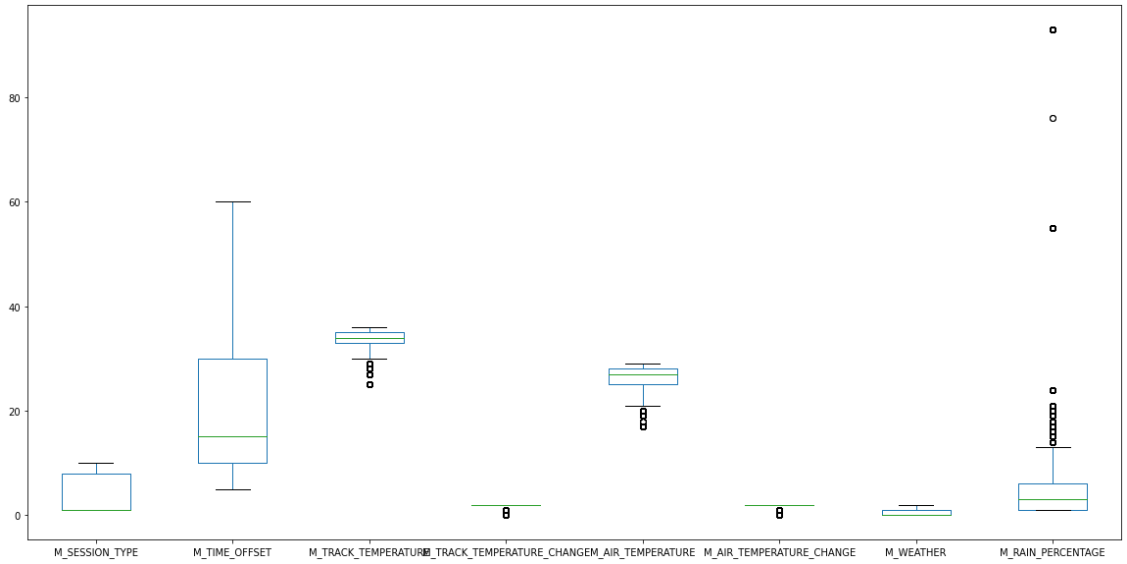
1) Neglecting certain columns as they were irrelevant or causing failures. For example, the M_ZONE_FLAG and M_ZONE_START had missing values alternating with the important features' missing values. Meaning that if we drop missing values on the entire dataset, results in an empty output.
2) Removing duplicate records as they provide no benefit to the model.
3) Filtering out values that have NUM_WEATHER_FORECAST_SAMPLES equals to 0, as stated by the competition tips as these records provide no value to the model.
4) Taking only the most relevant values into our data frame.
5) Removing missing records (NAN values).
6) Removing (0, 45, 90, 120) time offsets as our model needs to predict (5,10,15,30,60) offsets so they provide no benefit.

The dataset after cleaning and processing:

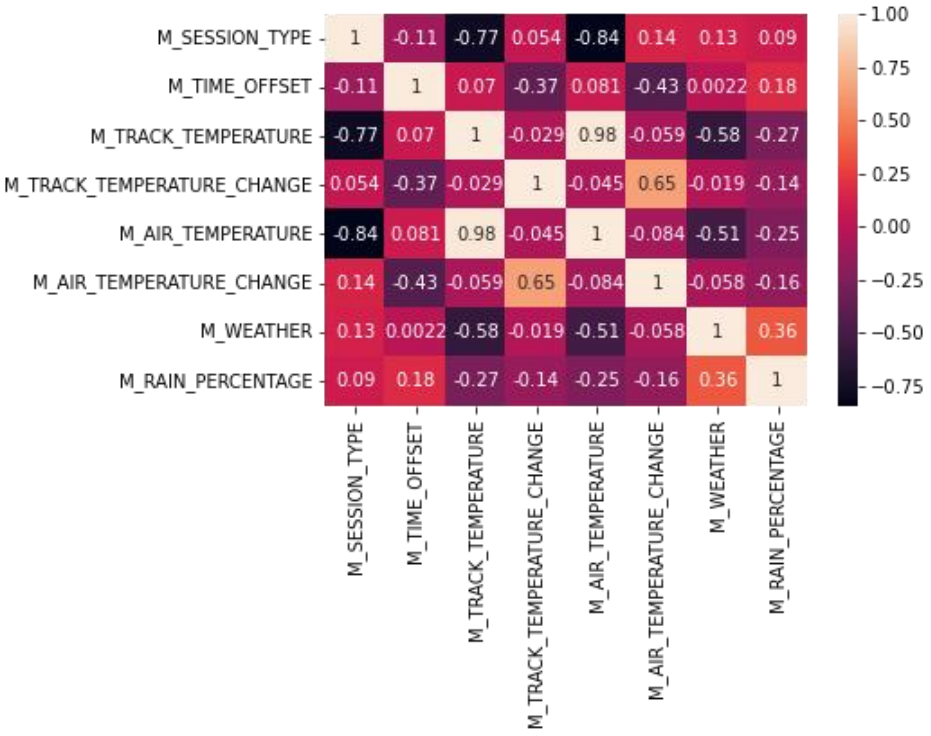
```
Data new shape : 457887 rows x 8 columns
```

	M_SESSION_TYPE	M_TIME_OFFSET	M_TRACK_TEMPERATURE	M_TRACK_TEMPERATURE_CHANGE	M_AIR_TEMPERATURE	M_AIR_TEMPERATURE_CHANGE	M_WEATHER	M_RAIN_PERCENTAGE
17039	8	5	33	2	25	2	0	1
17040	8	10	33	2	25	2	0	2
17042	8	5	33	2	25	2	0	5
17043	8	10	33	2	25	2	0	5
17044	8	15	33	2	25	2	0	7
...	...	...	...	...	...	...	...	...
3572286	8	5	33	2	25	2	0	3
3572287	8	10	33	2	25	2	0	3
3572288	8	15	33	2	25	2	0	3
3572289	8	30	33	2	25	2	0	3
3572291	8	60	33	1	25	2	0	3

Printing a **box plot**:



The **correlation matrix** after data cleansing:



Dataset statistics

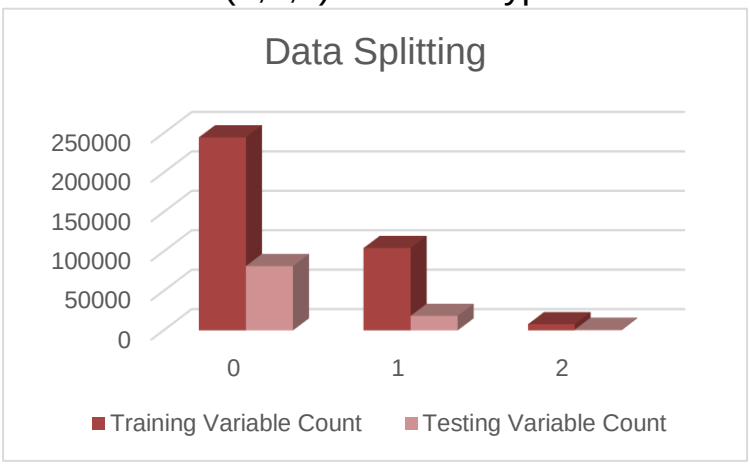
Number of variables	9
Number of observations	457887
Missing cells	0
Missing cells (%)	0.0%
Total size in memory	31.4 MiB
Average record size in memory	72.0 B

Splitting Data into Features and Labels

This splitting process is longer than usual as we have to ensure that both training and testing dataframes contain sufficient amount of records for (0,1,2) Weather types.

The process is as follows:

1) We split our data into 3 dataframes, one for each weather type (as we have 3 different weather types)
2) We split each of those frames into 2 more frames, one for the features and one for the labels. So, we end up having 6 dataframes



Builing up on the previous step, now we have to generate the training and testing data which can be achieved by:

1) Generating training and testing, features and labels for each dataframe of the previous step which results in 12 dataframe
2) We recombine those 12 dataframe into suitable frames that we desire, resulting in our main 4 dataframes (X_train, X_test, y_train, y_test)
3) Last step demonstrates how many value we end up with from each weather type value in both training and testing labels

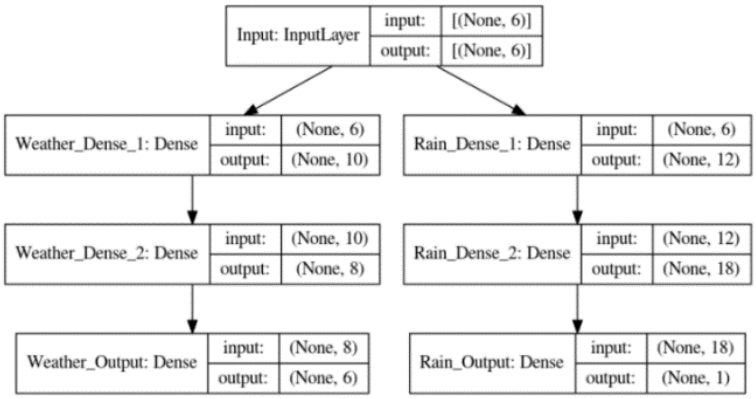
MODELS

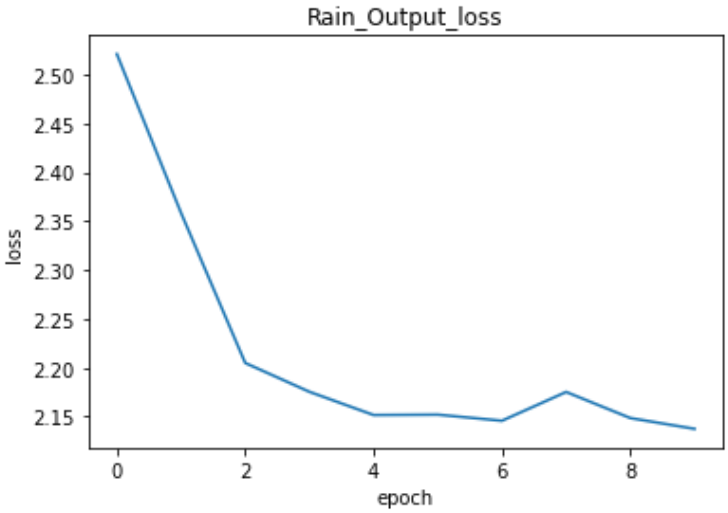
Model Selection

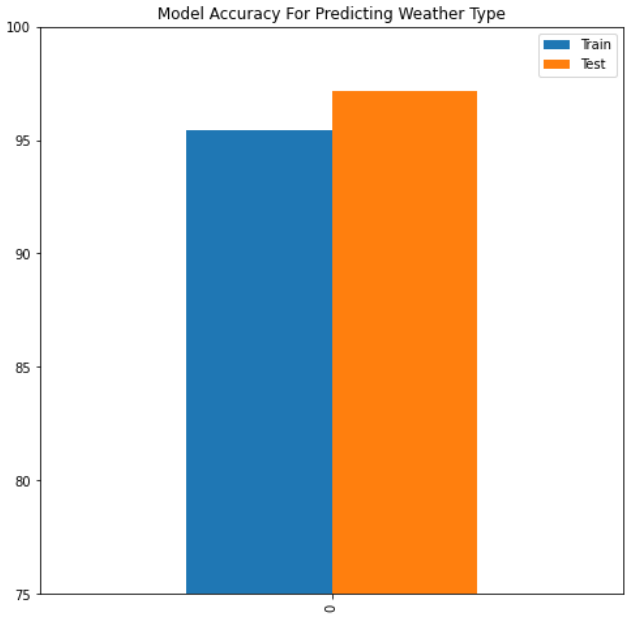
After testing each the of the following models, we chose to proceed with Multi-Output Neural Network.

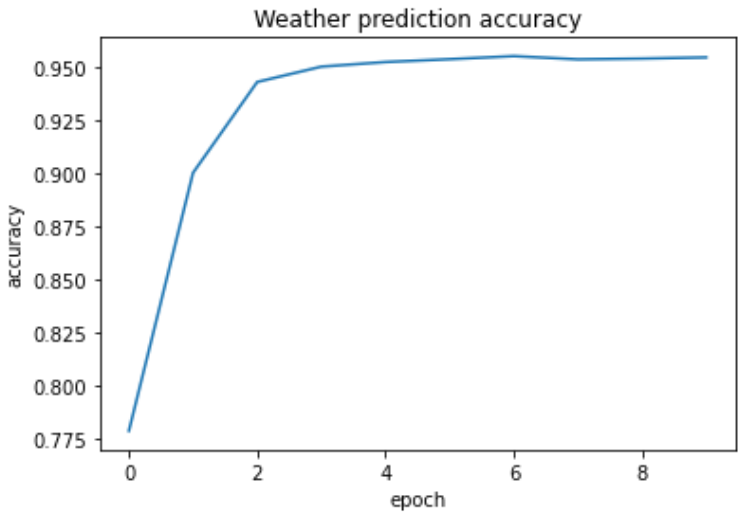
Models Implemented	Score
Decision Tree Regressor	MAE = 2.11
Decision Tree Classifier	ACC = 0.95
Logistic Regression	ACC = 0.96
Linear SVM	ACC = 0.93
Non Linear SVM	ACC = 0.96
Multi Output Neural Network	ACC = 0.97 MAE = 1.8

The hierarchy of the chosen Branched Neural Network.









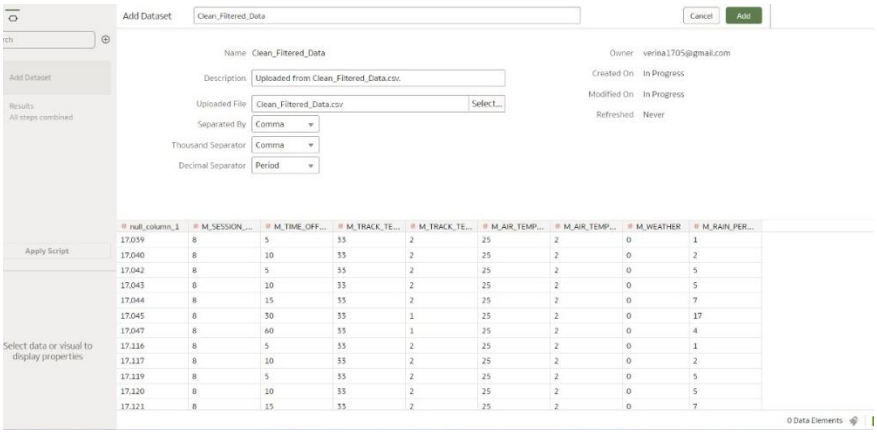
Output Dictionary

After setting the format of the output to the required format, the following was achieved as a predicted sample.

```
{'5': {'Type': 0, 0: 4.956110954284668},
 '10': {'Type': 0, 0: 4.950217247009277},
 '15': {'Type': 0, 0: 4.863468170166016},
 '30': {'Type': 0, 0: 1.8930506706237793},
 '60': {'Type': 0, 0: 9.26978874206543}}
```

Integration with Oracle

first, we analyzed the data on Oracle Dashboard for data analysis and then we analyzed the filtered data. A comparison was made to validate our work.



Then Oracle Cloud was used to run our notebook, also to validate the model results.