New York, JFK Airport Flight Data

Linear Regression

the DEVELOPER Academy

by Kieran Thakkar

About Dataset

Context

This data was scraped under a Academic Paper under Review by IEEE transportation

Content

This file contains data about flights leaving from JKF airport between Nov 2019-Dec-2020.

Taxi-Out

Taxi-Out prediction is an important concept as it helps in calculating Runway time and directly impact the cost of the flight.

Data:

Date of the flight (month, day of the month, day of the week),

information about the operating carrier (OP_UNIQUE_CARRIER),

the tail number of the plane (TAIL_NUM),

the destination of the flight (DEST),

the delay of the departure (DEP_DELAY),

the elapsed time of the flight (CRS_ELAPSED_TIME),

the distance traveled (DISTANCE),

the scheduled and actual departure time (CRS_DEP_M, DEP_TIME_M),

the scheduled and actual arrival time (CRS_ARR_M),

weather conditions (Temperature, Dew Point, Humidity, Wind, Wind Speed, Wind Gust, Pressure, Condition),

the scheduled departure and arrival times (sch_dep, sch_arr),

and the taxi-out time (TAXI_OUT).

Data:

Date of the flight (month, day of the month, day of the week),

~~information about the operating carrier (OP_UNIQUE_CARRIER),~~

~~the tail number of the plane (TAIL_NUM),~~

~~the destination of the flight (DEST),~~

the delay of the departure (DEP_DELAY),

the elapsed time of the flight (CRS_ELAPSED_TIME),

the distance traveled (DISTANCE),

the scheduled and actual departure time (CRS_DEP_M, DEP_TIME_M),

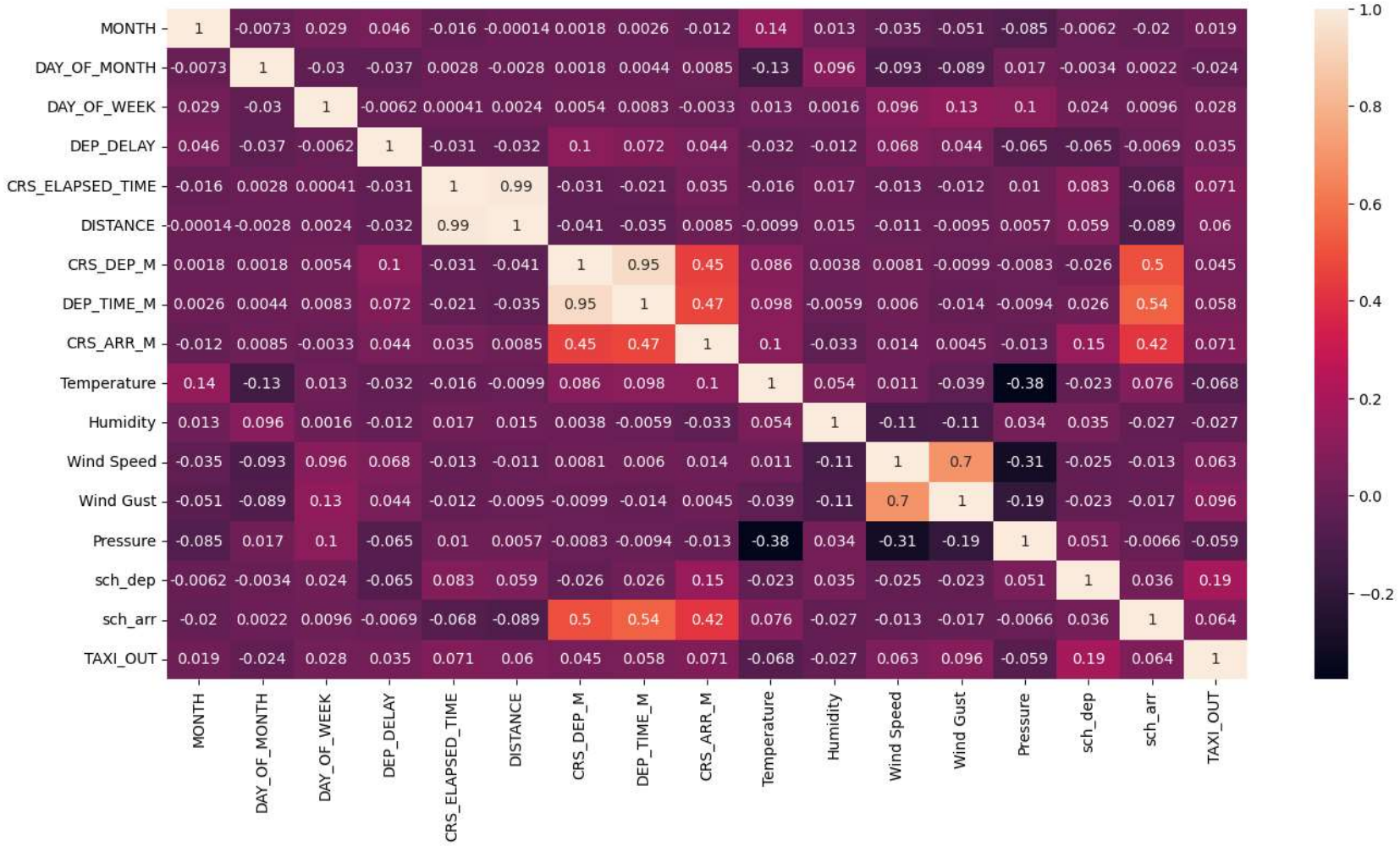
the scheduled and actual arrival time (CRS_ARR_M),

weather conditions (Temperature, ~~Dew Point~~, Humidity, ~~Wind~~, Wind Speed, Wind Gust, Pressure, ~~Condition~~),

the scheduled departure and arrival times (sch_dep, sch_arr),

and the taxi-out time (TAXI_OUT).

Correlation Matrix



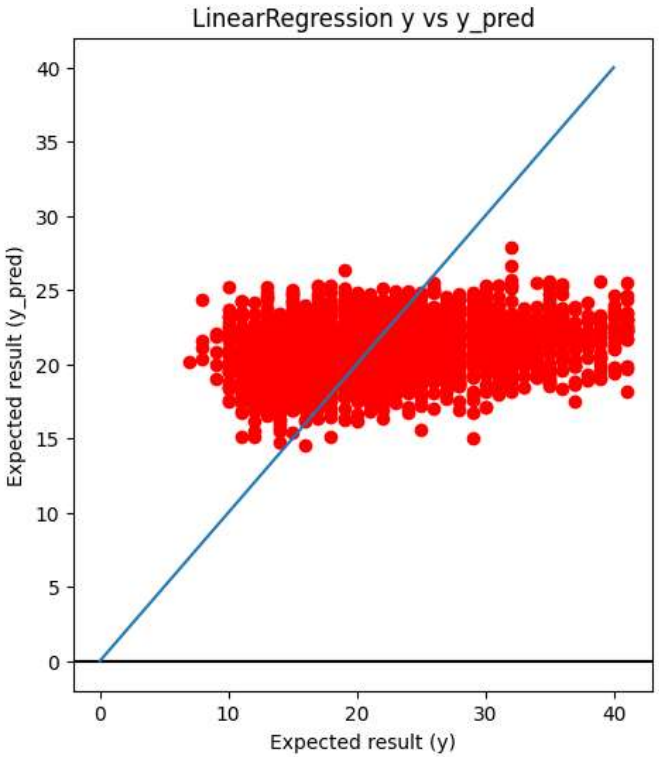
Numeric data only.

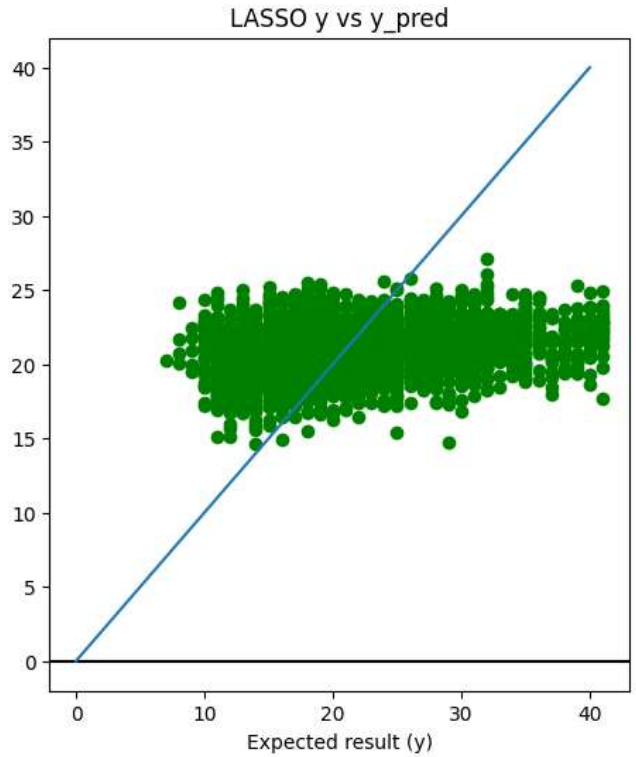
Split data into 3 parts

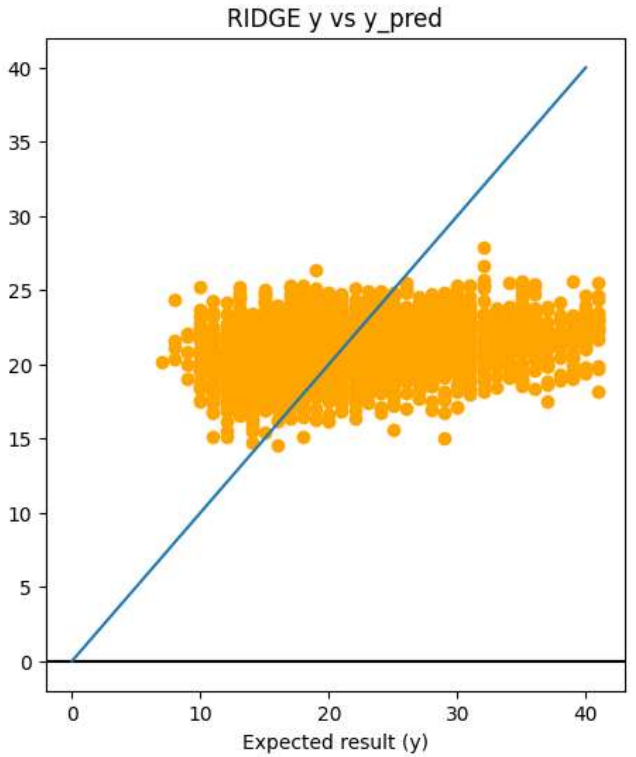
- All numeric data
- Weather data
- Scheduled departure time

All numeric data

Using every numeric column to predict Taxi-Out time.
Scatter plot should fall on this diagonal line where y = y_predicted.







All data: Regression metrics

```
LinearRegression()
MAE: 5.294266838975587
MSE: 44.87203650888366
RMSE: 6.698659306822796
R2: 0.07908317548672061
```

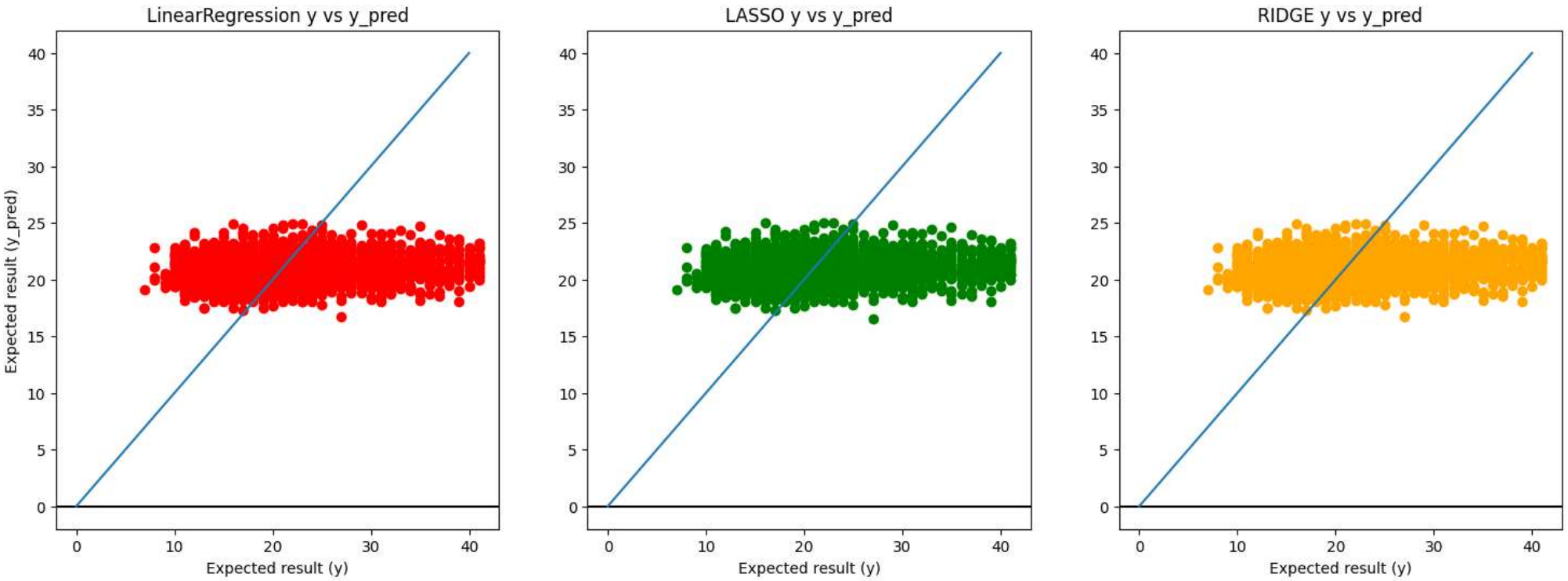
```
Lasso(alpha=0.1)
MAE: 5.308590161010935
MSE: 45.16678082276287
RMSE: 6.720623544193118
R2: 0.07303408525817146
```

```
Ridge(alpha=1)
MAE: 5.294265206463216
MSE: 44.87209430731264
RMSE: 6.698663621000285
R2: 0.07908198927923327
```

Weather data

Using only weather data to predict Taxi-Out time.

Scatter plot should fall on this diagonal line where y = y_predicted.



Weather data: Regression metrics

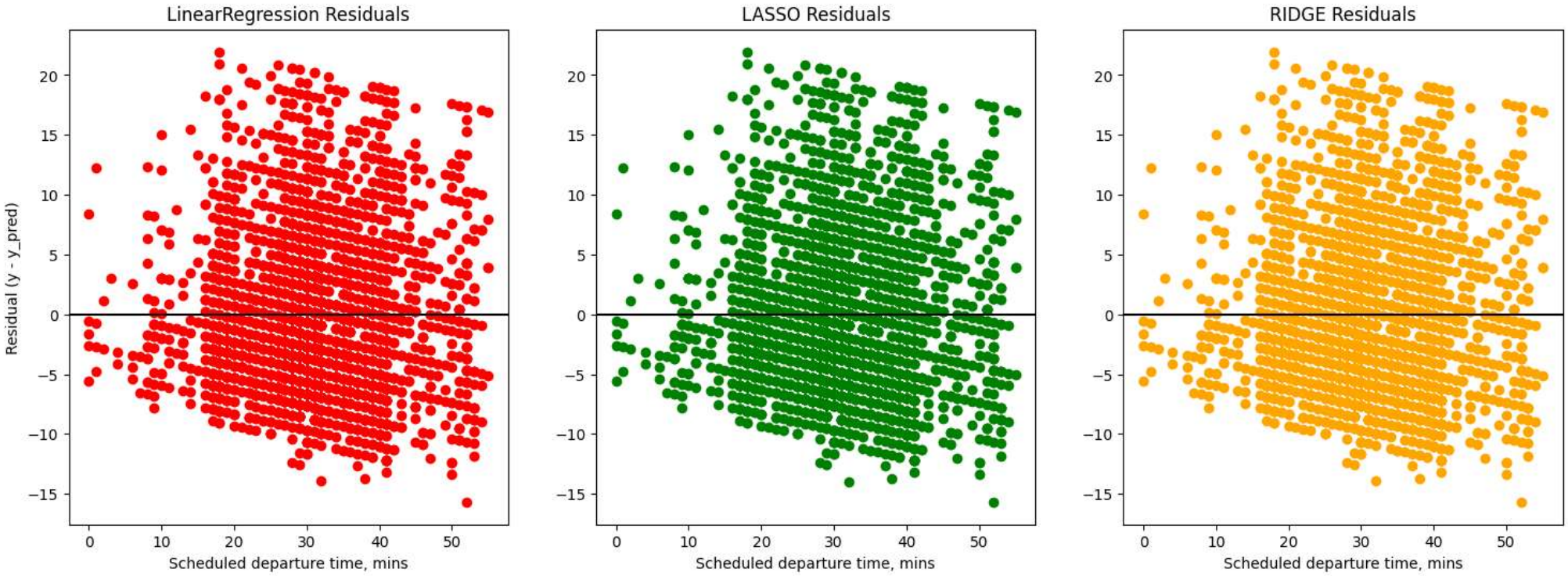
```
LinearRegression()
MAE: 5.4797779872400225
MSE: 47.268086585745515
RMSE: 6.875179022087026
R2: 0.029908611552641795
```

```
Lasso(alpha=0.1)
MAE: 5.308590161010935
MSE: 45.16678082276287
RMSE: 6.720623544193118
R2: 0.07303408525817146
```

```
Ridge(alpha=1)
MAE: 5.294265206463216
MSE: 44.87209430731264
RMSE: 6.698663621000285
R2: 0.07908198927923327
```

Scheduled departure time data

Using only sch_dep column to predict Taxi-Out time.
Data points plotted should be close to 0. Not the case.



Weather data: Regression metrics

```
LinearRegression()
MAE: 5.448049298058799
MSE: 46.928190459177436
RMSE: 6.85041534939141
R2: 0.036884360502245 86
```

```
Lasso(alpha=0.1)
MAE: 5.448106462548827
MSE: 46.928343137732824
RMSE: 6.850426493126747
R2: 0.036881227052967924
```

```
Ridge(alpha=1)
MAE: 5.448049298354524
MSE: 46.928190459973386
RMSE: 6.850415349449505
R2: 0.036884360485910 37
```

Conclusions

- Don’t waste your time trying to find a correlation

- Best relationship found using all the data together w/ `LinearRegression()`.
- Very little correlation, R^2 = `0.07908317548672061`

Improvements

- Group data by wind direction
- Group data by weather conditions
- Normalise temperature data relative to standard (~20°C)

Thanks for watching !