JFK TAXI-OUT Linear Regression

Nov 2019 – Jan 2020

Deirdre Boland
07 Feb 2025

Objective and dataset

Objective

- Create Linear Regression Model of JFK Taxi-Out, using data scraped from an Academic Paper under Review by IEEE transportation covering Nov 2019- Jan 2020 (D Kansal, Kaggle dataset)

Value

- At JFK airport Taxi-Out prediction is an important concept for calculating runway time and directly impacts the cost of flights.

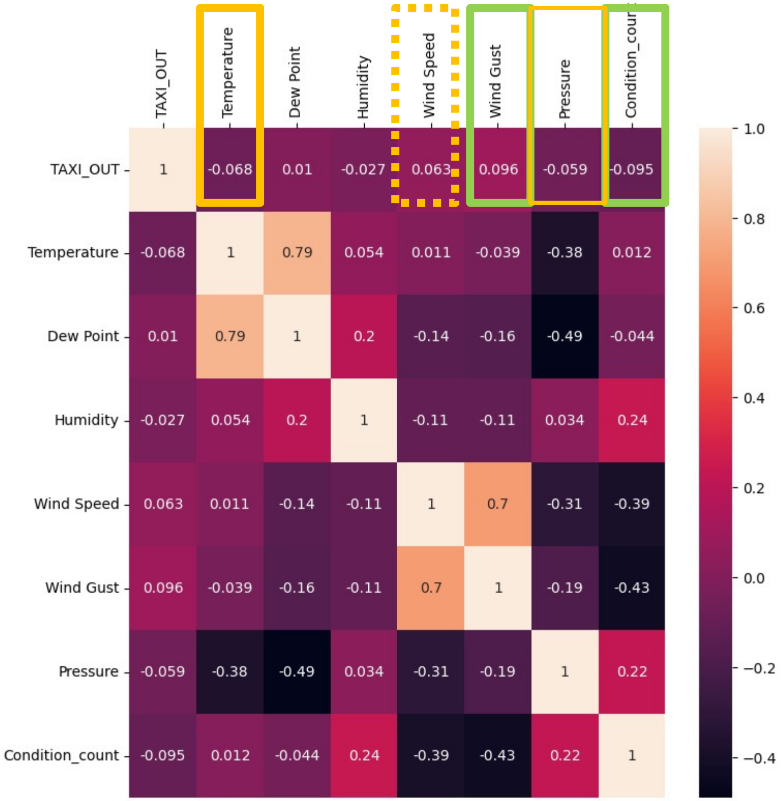
Dataset

- 5 text based features/variables
  - Airline and flight number indicators (TAIL_NUM, OP_UNIQUE_CARRIER)
  - Destination (DEST) covered by numeric features of distance and scheduled flight time
  - Wind direction (Wind e.g. NW, E etc) – have wind speed and gust in numeric
  - Climate/weather (Condition) – converted to numeric by using frequency counts (more severe weather less frequent)
- 18 numeric features, including target feature(TAXI_OUT)
  - Covering weather/conditions at flight time and flight details e.g. Time, duration
  - Scheduled Arrival Time (CRS_ARR_M) excluded – arrival at another airport, includes time difference offsets. This feature is covered by scheduled duration of flight feature
- Final 18 numeric features (17 original & 1 generated from Condition text feature)

- Reference – Kaggle page Flight Take Off Data - JFK Airport

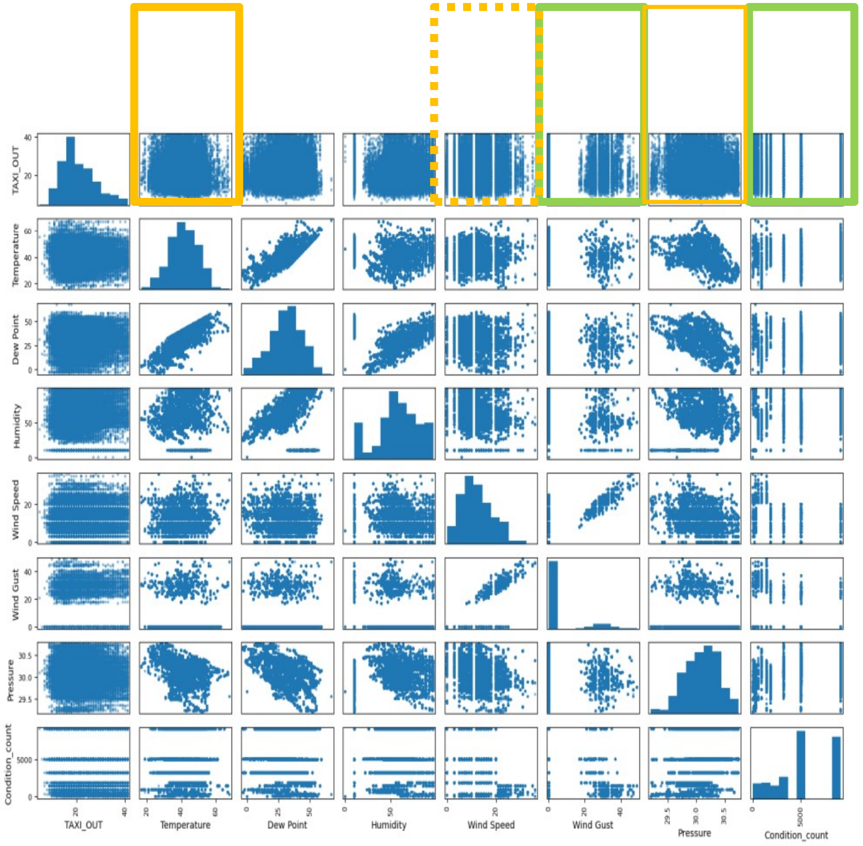
Correlation Matrix – Weather/Climate

Correlation Matrix



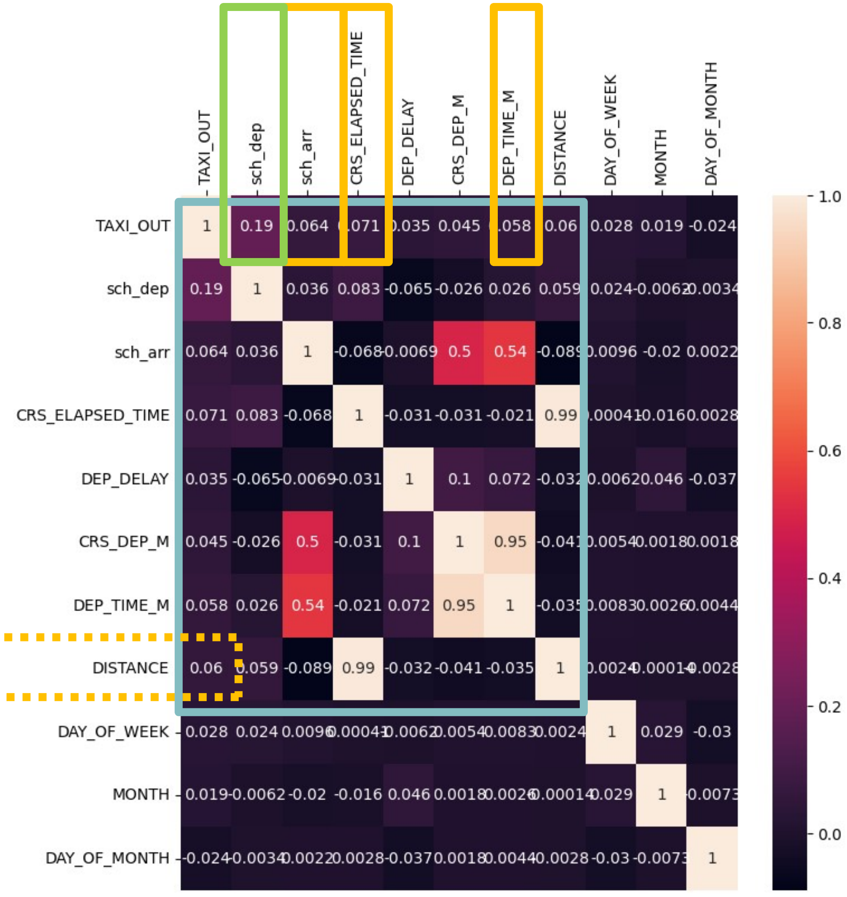
	TAXI_OUT	Temperature	Dew Point	Humidity	Wind Speed	Wind Gust	Pressure	Condition_cour
TAXI_OUT	1	-0.068	0.01	-0.027	0.063	0.096	-0.059	-0.095
Temperature	-0.068	1	0.79	0.054	0.011	-0.039	-0.38	0.012
Dew Point	0.01	0.79	1	0.2	-0.14	-0.16	-0.49	-0.044
Humidity	-0.027	0.054	0.2	1	-0.11	-0.11	0.034	0.24
Wind Speed	0.063	0.011	-0.14	-0.11	1	0.7	-0.31	-0.39
Wind Gust	0.096	-0.039	-0.16	-0.11	0.7	1	-0.19	-0.43
Pressure	-0.059	-0.38	-0.49	0.034	-0.31	-0.19	1	0.22
Condition_count	-0.095	0.012	-0.044	0.24	-0.39	-0.43	0.22	1

Scatter Matrix



Correlation Matrix – Flight Details

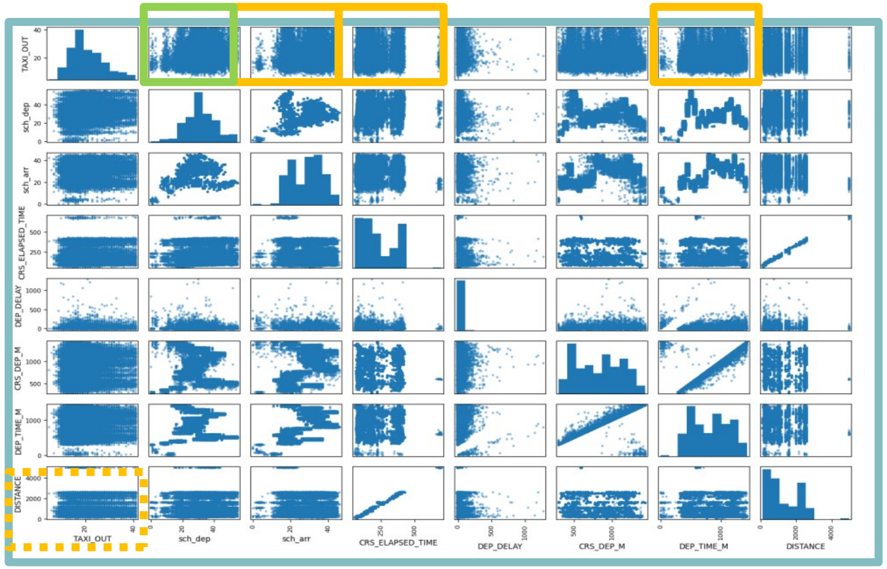
Correlation Matrix



	TAXI_OUT	sch_dep	sch_arr	CRS_ELAPSED_TIME	DEP_DELAY	CRS_DEP_M	DEP_TIME_M	DISTANCE	DAY_OF_WEEK	MONTH	DAY_OF_MONTH
TAXI_OUT	1	0.19	0.064	0.071	0.035	0.045	0.058	0.06	0.028	0.019	-0.024
sch_dep	0.19	1	0.036	0.083	-0.065	-0.026	0.026	0.059	0.024	-0.0062	0.0034
sch_arr	0.064	0.036	1	-0.068	-0.0069	0.5	0.54	-0.089	0.0096	-0.02	0.0022
CRS_ELAPSED_TIME	0.071	0.083	-0.068	1	-0.031	-0.031	-0.021	0.99	0.00041	-0.016	0.0028
DEP_DELAY	0.035	-0.065	-0.0069	-0.031	1	0.1	0.072	-0.032	0.0062	0.046	-0.037
CRS_DEP_M	0.045	-0.026	0.5	-0.031	0.1	1	0.95	-0.041	0.0054	0.0018	0.0018
DEP_TIME_M	0.058	0.026	0.54	-0.021	0.072	0.95	1	-0.035	0.0083	0.0026	0.0044
DISTANCE	0.06	0.059	-0.089	0.99	-0.032	-0.041	-0.035	1	0.0024	0.00014	0.0028
DAY_OF_WEEK	0.028	0.024	0.0096	0.00041	0.0062	0.0054	0.0083	0.0024	1	0.029	-0.03
MONTH	0.019	-0.0062	-0.02	-0.016	0.046	0.0018	0.0026	0.00014	0.029	1	-0.0073
DAY_OF_MONTH	-0.024	-0.0034	0.0022	0.0028	-0.037	0.0018	0.0044	0.0028	-0.03	-0.0073	1

Scatter Matrix

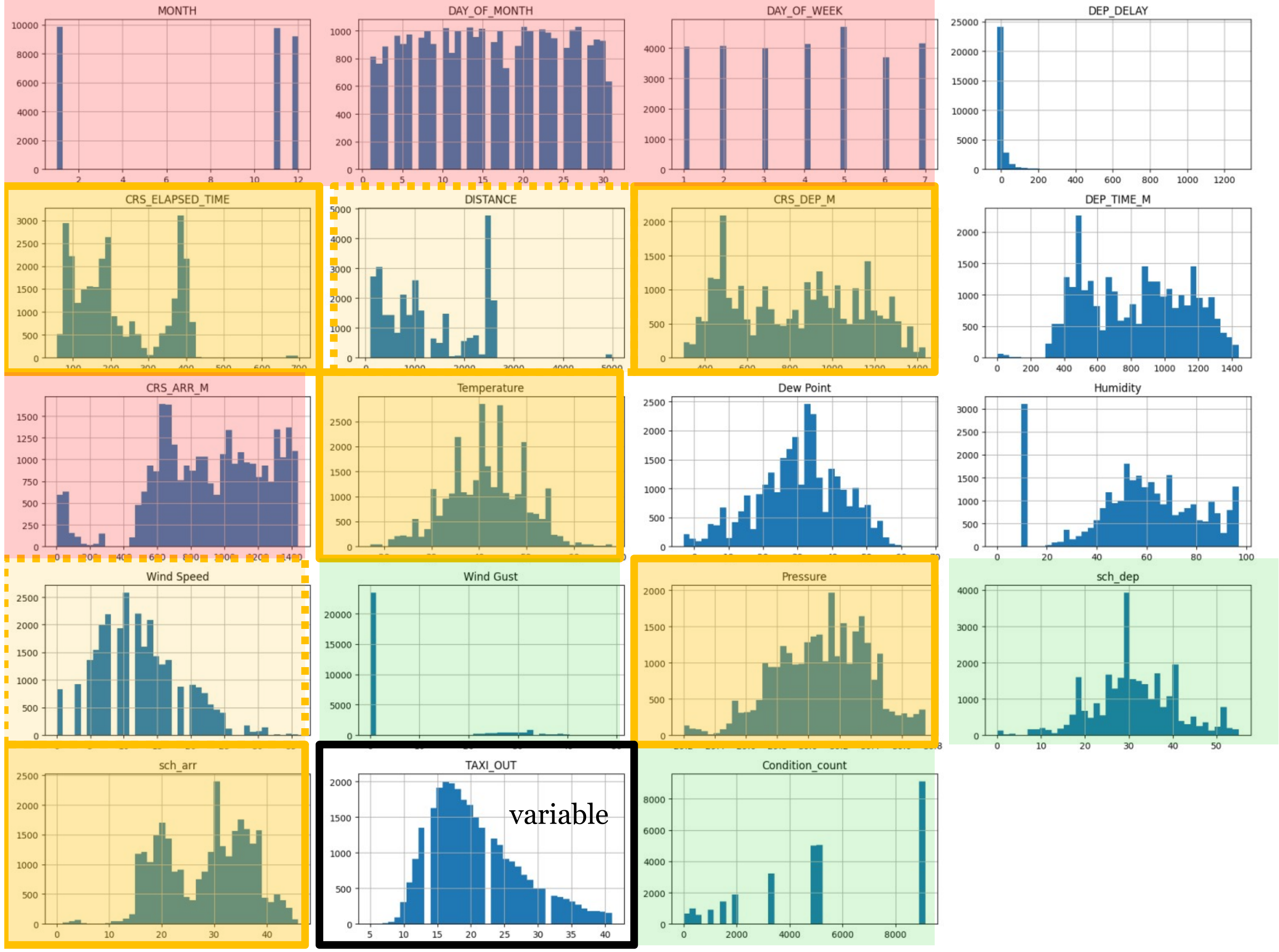
- Number of flights scheduled for arrival./departure (sch_arr, sch_dep)
- Departure delay of the flight (DEP_DELAY) - is calculation of *Actual Departure Time (DEP_TIME_M) subtract Scheduled Departure Time (CRS_DEP_M)
  * Gate checkout of the flight not the take off time



- Scheduled journey time of the flight (CRS_ELAPSED_TIME)

Linear regression Model

Feature Histograms
(x- axis, horiz = units bin 40, y- axis, vert = frequency)



- *All models scaled using standard scaler. 80% train, 20% test*
- R² 0-1, higher is better fit

3 features (highest corrl, green)

- Multi Linear Regression (MLR) or Ordinary Least Squares (OLS) – r² 0.059
- Lasso – r² 0.057
- Ridge – r² 0.059

8 features (>0.5 corrl, ambr non-dash & green)

- MLR/OLS – r² 0.077
- Lasso – r² 0.074
- Ridge – r² 0.077

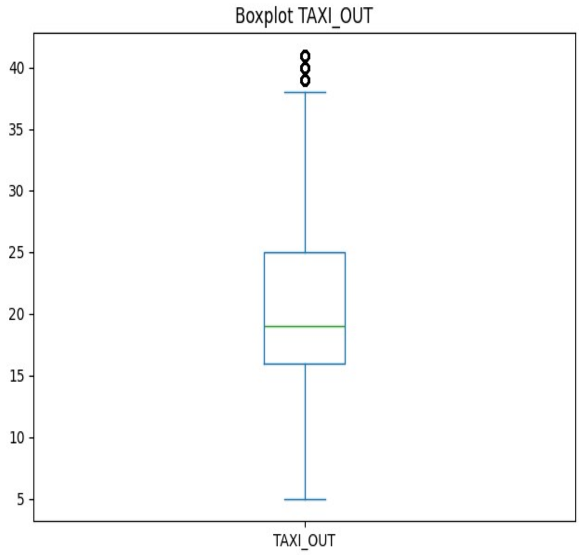
14 features (all except red)

- MLR/OLS – r² 0.083
- Lasso – r² 0.080
- Ridge – r² 0.083

Alpha tuning (set at 0.1 above)

- *Regularisation penalty in ridge and lasso modelling.*
- *Increasing value on Lasso made model dramatically worse but minimal impact Ridge*
- *Expected as increasing alpha will get rid of features on Lasso*

Conclusion and next steps



- Linear regression model not viable to model JFK Taxi-out with this data set
  - low correlation of features leading to underfitting (erroneous outcomes on new data)
- Limited dataset Nov 2019 – Jan 2020, includes holidays and wintry conditions
- Majority Taxi-out time under 25 mins with a min of 5 and max of 40mins
  - Need more information on Taxi-out metric and calculation
    - What is target? Is consistent 5-15 min realistic?
    - What other data not collected, might be impacting? E.g. impact of different terminals (5) and runways (4), staffing levels

Q&A