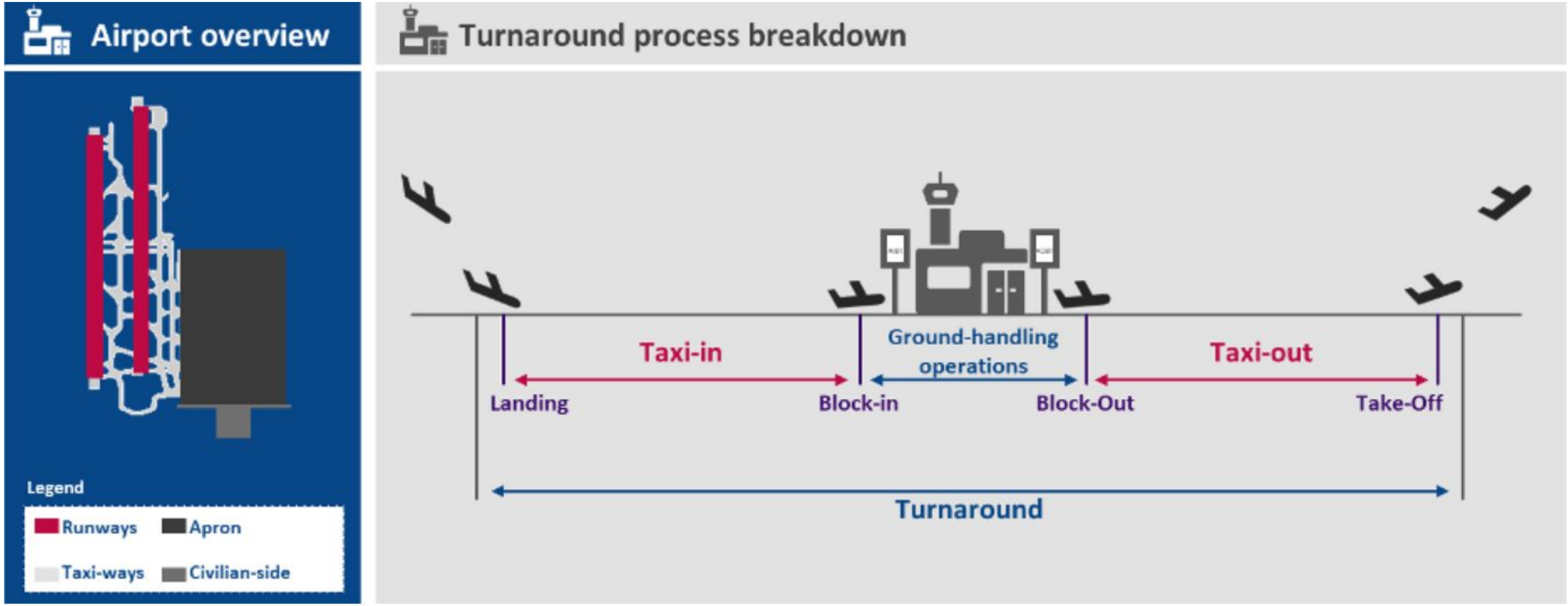
What is the best predictor for taxi out time?

Content

How does an airport work?



Reducing the taxi-out time

Taxi-out: **timespan between block-out and take-off**

It affects:

- Airline companies
- Airports
- Environment

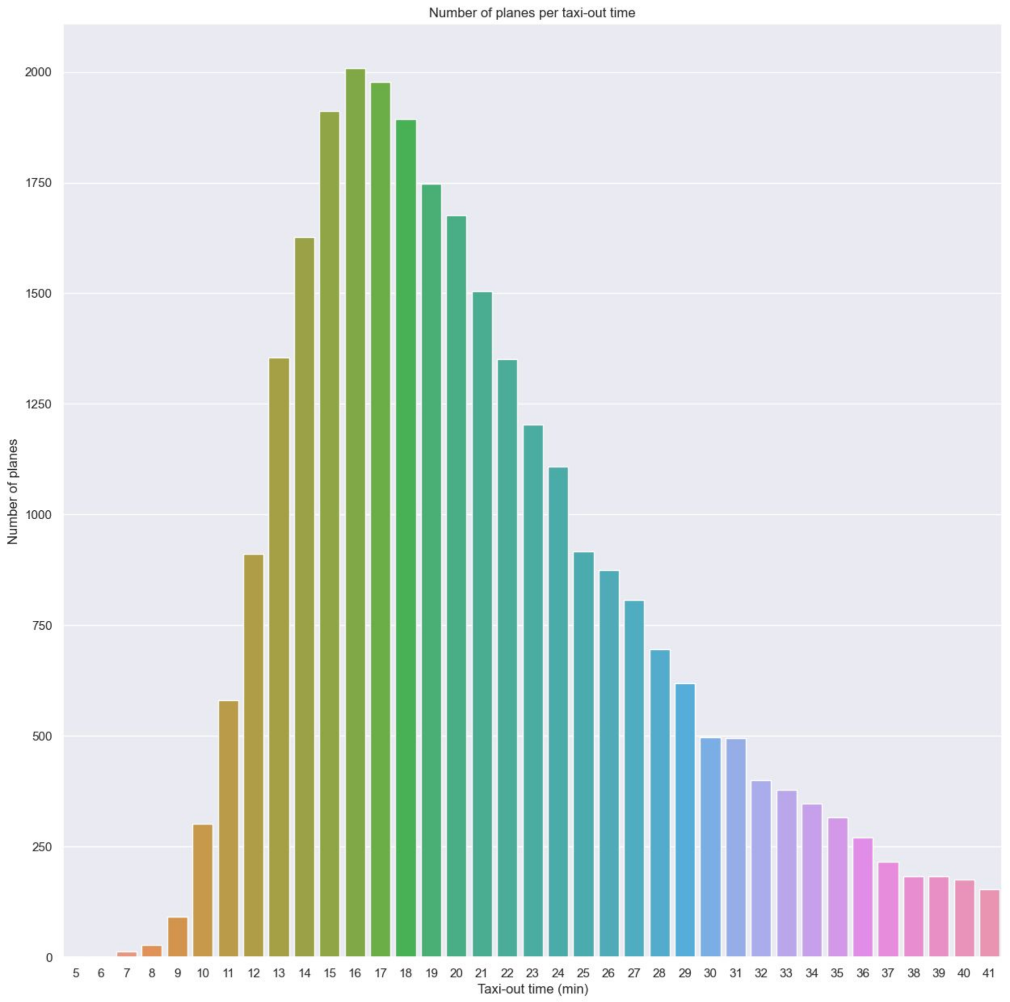
Solution: Using linear regression to find the best predictor

The dataset

Data about flights leaving JFK Airport (Nov 2019 - Dec 2020)

- 28818 entries after removing 2
- 23 variables
- Target variable: taxi_out

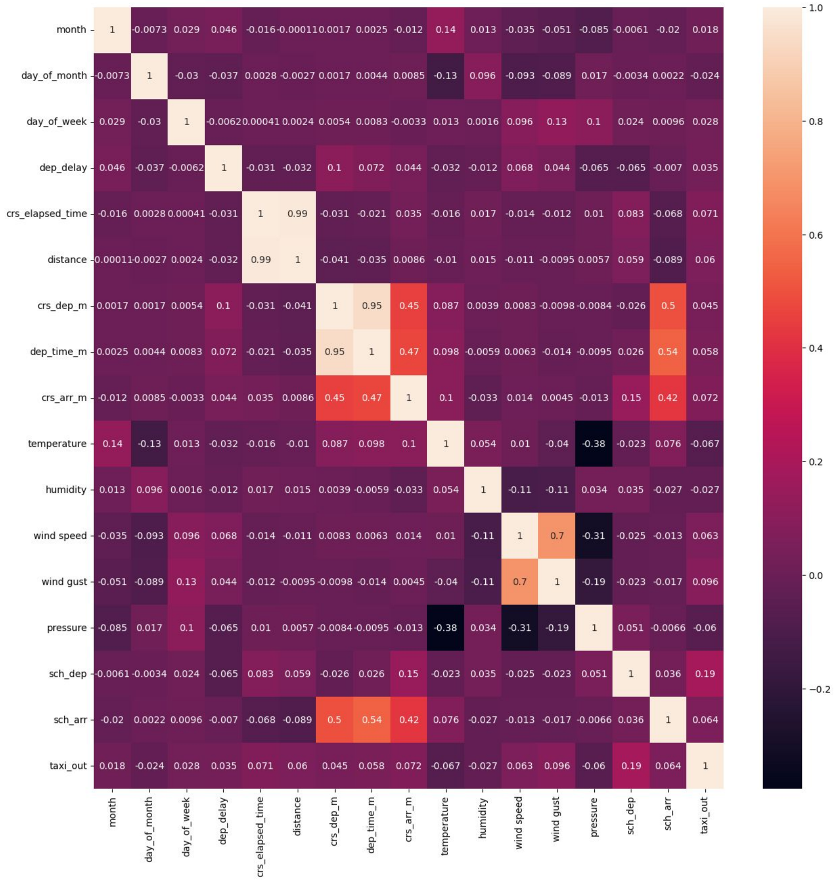
```
Index: 28818 entries, 0 to 28819
Data columns (total 23 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   month              28818 non-null  int64
 1   day_of_month       28818 non-null  int64
 2   day_of_week        28818 non-null  int64
 3   op_unique_carrier  28818 non-null  object
 4   tail_num           28818 non-null  object
 5   dest               28818 non-null  object
 6   dep_delay          28818 non-null  int64
 7   crs_elapsed_time   28818 non-null  int64
 8   distance           28818 non-null  int64
 9   crs_dep_m          28818 non-null  int64
 10  dep_time_m         28818 non-null  int64
 11  crs_arr_m          28818 non-null  int64
 12  temperature        28818 non-null  int64
 13  dew point          28818 non-null  object
 14  humidity           28818 non-null  int64
 15  wind               28818 non-null  object
 16  wind speed         28818 non-null  int64
 17  wind gust          28818 non-null  int64
 18  pressure           28818 non-null  float64
 19  condition          28818 non-null  object
...
 21  sch_arr            28818 non-null  int64
 22  taxi_out           28818 non-null  int64
dtypes: float64(1), int64(16), object(6)
memory usage: 5.3+ MB
```



Frequency bar plot

- Most frequent : 16min
- Mean: 20.85 min
- Min: 5
- Max: 41

Considered delayed if over 15 min



Correlation Matrix

- All below 0.1
- **Except** for sch_dep (number of flights scheduled for departure) at 0.19

Next top variables:

- Wind gust: 0.096
- Crs_elapsed_time (schedule journey time of the flight): 0.071

Evaluation of the model

Linear Regression with number of flights scheduled on for departure

```
R squared: 0.04
Mean Absolute Error:  5.353282616449877
Mean Square Error:  45.10840350858983
Root Mean Square Error:  6.716278992760041
```

Multiple Regression with the 3 highest correlated variables

```
R squared: 0.05
Mean Absolute Error:  5.318760437051704
Mean Square Error:  44.56825084333637
Root Mean Square Error:  6.675945689064312
```

Best model

Multiple Regression with the all numerical variables

```
R squared: 0.07
Mean Absolute Error:  5.260479190557299
Mean Square Error:  43.62204291226953
Root Mean Square Error:  6.604698548175346
```

Conclusion

Number of flights scheduled for departure is the best predictor here.

But not the only one.

Further investigation:

- Clean more data
- More exploration: time series
- Feature engineer : encoding, etc
- Explore categorical data
- Use other models

—> Optimise time-out prediction —> £££


THANK YOU!

Appendix

How does an airport work?

1. **ALDT :** Actual Landing Time - wheels on the ground
2. **AIBT :** Actual In-Block Time - reach dock at registered timestamp
3. Ground-handling teams (catering, cleaning, fueling, boarding…)
4. **AOBT :** Actual Off-Block Time - left dock
5. **ATOT :** Actual Take-Off Time - wheels off the ground

The issues with taxi-out

For airline companies: the basics of airline companies' finances in airports is that an aircraft that is not airborne is an aircraft losing money. So, by providing a better taxi-out time prediction and reducing queues at the runway's entry point, the solution ensures less time spent on the ground for A/Cs, and thus less money lost for airline companies
For airports: for an airport, the most A/Cs are operated per day, the more money it makes. So by providing a forecast that smoothens the A/C flow, the taxi-out time prediction solution potentially increases the amount of aircrafts that can be operated per day at the airport and thus the money generated by the airport
For the environment: when an aircraft is queuing at the runway's entry point, one should know that its engines are still running, so that's kerosene that is used and GHG emissions generated for virtually nothing. By reducing the queuing time, the solution also reduces the GHG emission levels at airports.

https://leonard.vinci.com/en/taxi-out-time-prediction/

Dataset

- **Table 1.** Attribute description for the data set. :

https://dl.acm.org/doi/fullHtml/10.1145/3497701.3497725

- From **Kaggle**:

https://www.kaggle.com/deepankurk/flight-take-off-data-jfk-airport/tasks?taskId=4868

Github link to code of models

Made & presented on : 02/06/2023