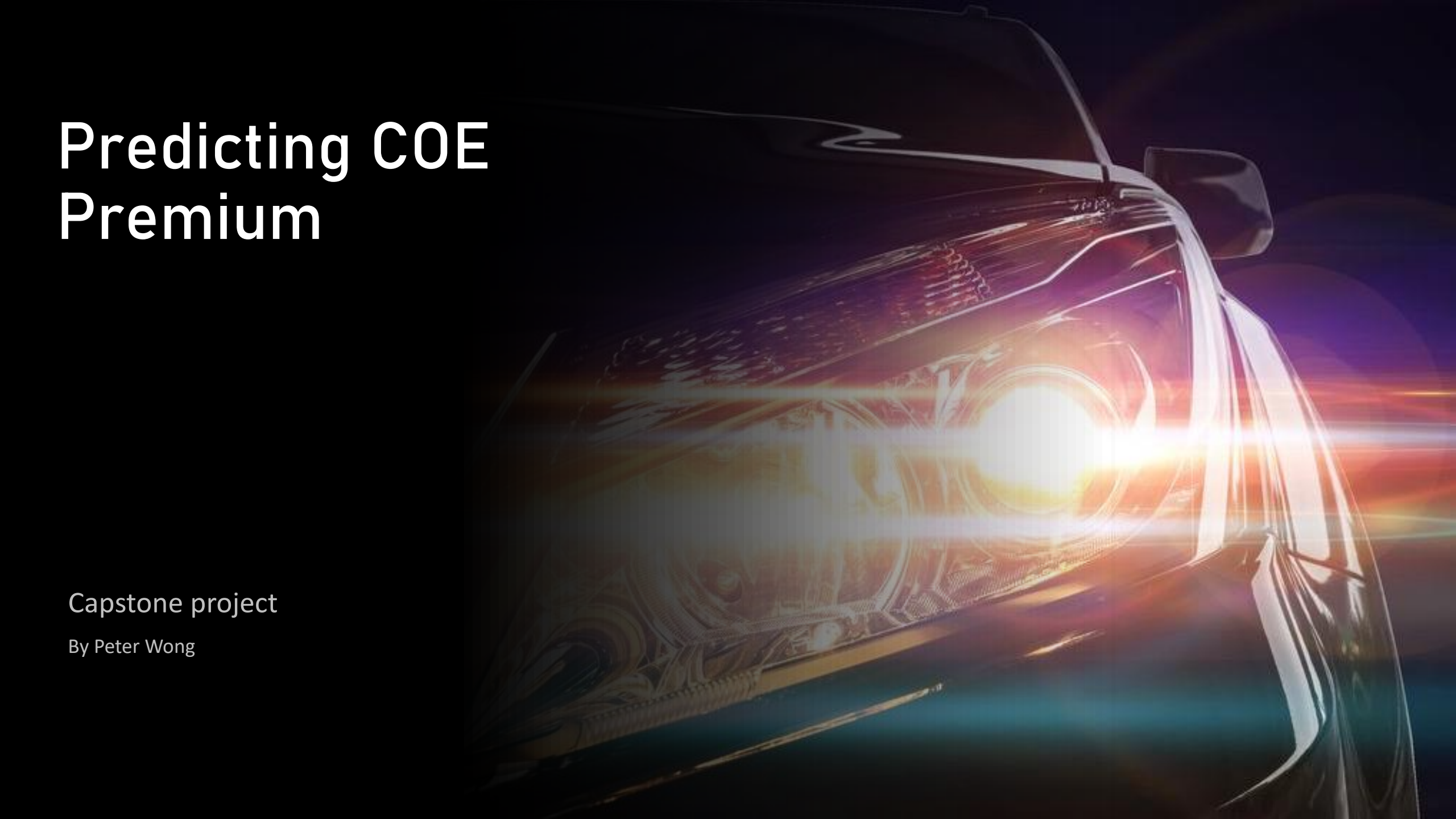
# Predicting COE Premium



Capstone project

By Peter Wong

## Agenda

- Problem Statement
- Background
- Data
- Modelling
- Conclusions
- Demo

# Problem Statement

With more retailers and shoppers moving online, there is an increased demand for delivery services. More corporations are looking to expand their delivery and private hire fleet. This is a pilot project to predict the COE premium for budgetary purposes. With a good prediction, stakeholders would be in a better position to plan and allocate budgets.

Would classic time series or regression models be able to achieve this with a RMSE of 5K or less.

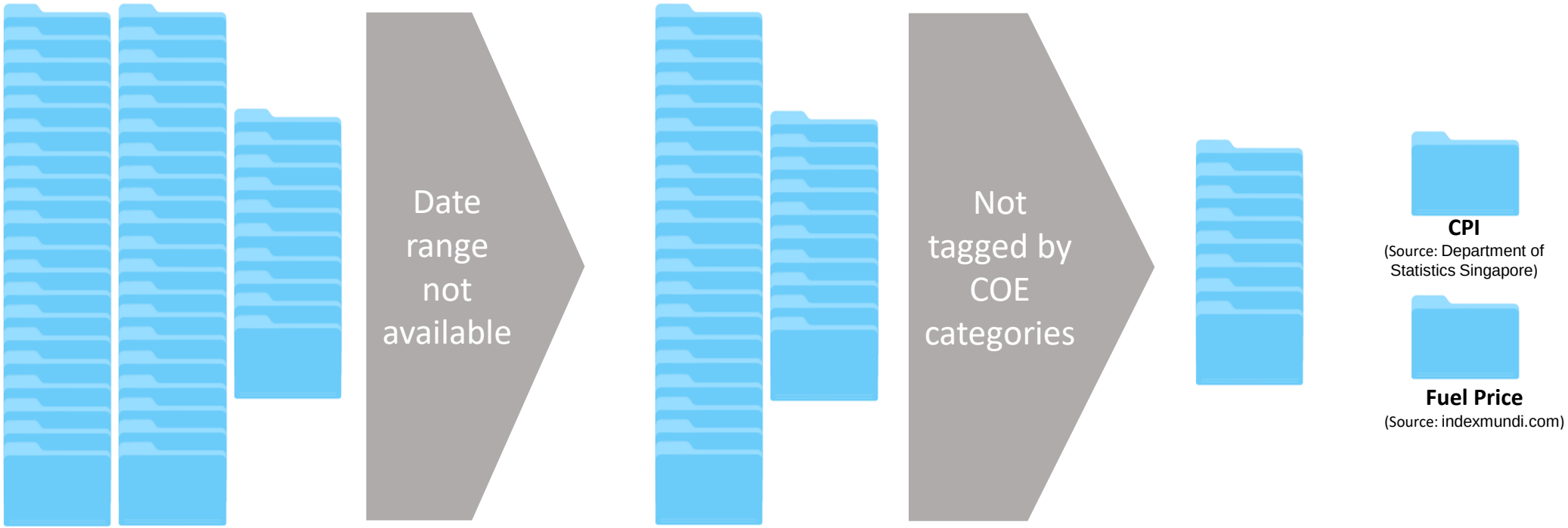
**RMSE of 5K or less**

# Background

All vehicles in Singapore require a COE. To register a vehicle, you must first place a bid for a Certificate of Entitlement (COE) in the corresponding vehicle category. A successful COE bid gives you the right to own a vehicle that can be used on the road for 10 years.
COEs are released through open bidding exercises. At the end of the 10-year COE period, you can choose to deregister your vehicle, or renew your COE.


**Project focus: Category A, B and C**

# Data




46 files
(source: LTA)

25 files

6 files

**Features**

['vehicle_class', 'quota', 'bids_success', 'bids_received', 'premium', 'cpi', 'fuel_price', 'reg', 'dereg']

**Date range: 2010 to 2020, Features count 9, row count 378**

# Model Selection

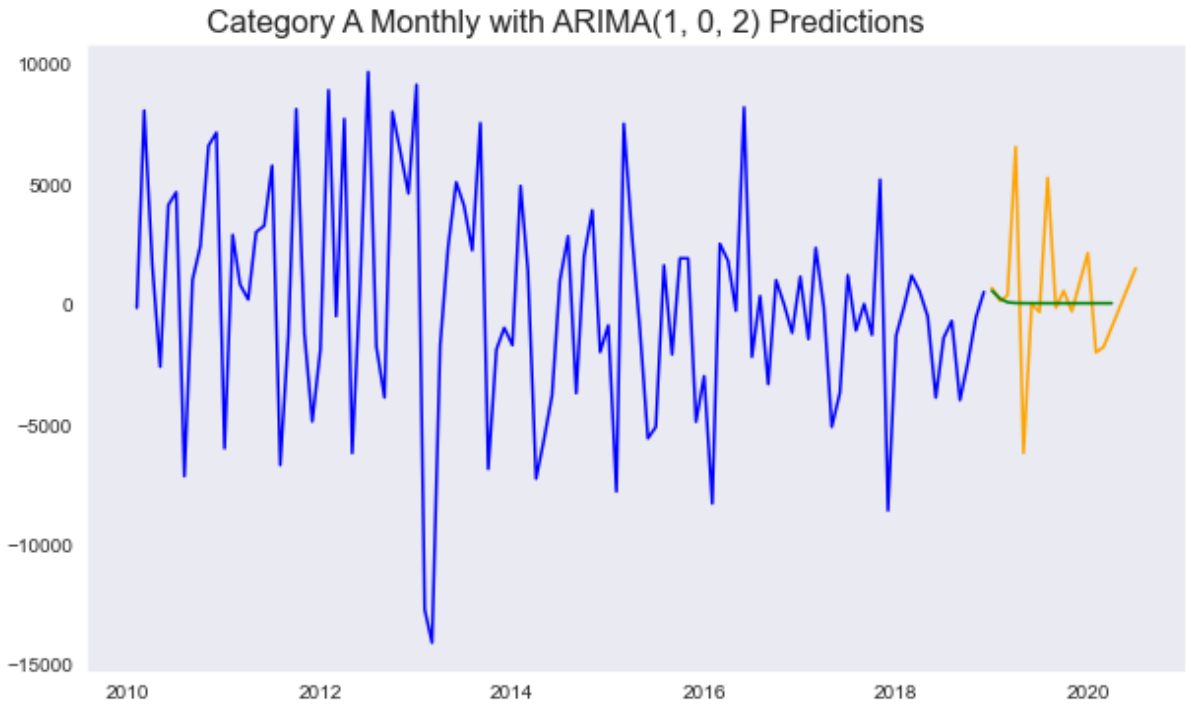
# Modelling ARIMA

Classic Time series model, **ARIMA**

Target variable: **Premium** (not stationary)

Took diff order 1 of Premium, to GridSearch for best ARIMA order.

Predict with best ARIMA order, but this didn’t yield good results.




**ARIMA didn’t yield very good results**

# Modelling + Features Engineering

| Shifted/lagged features |
| --- |
| 'cpi', 'fuel_price', 'dereg', 'premium' |
| 1, 2, 3, 5, 10, 15, 20, 24, 48 month |

| Exponentially weighted moving average |
| --- |
| 'bids_success', 'bids_received', 'quota' |
| EWMA 3 months |

**Linear Regression**

**XGBoost Regressor**

| | | |
| --- | --- | --- |
| **Initial RMSE** | **~18K average across category** | **~12K average across category** |

**Created shifted and EWMA features, Features count 45**

# Model Metrics

| | Category A | | Category B | | Category C | |
|---|---|---|---|---|---|---|
| | **lrA** | **xgbA** | **lrB** | **xgbB** | **lrC** | **xgbC** |
| **01 Train R2** | 0.9351 | 0.9989 | 0.9225 | 1 | 0.8619 | 0.9991 |
| **02 Test R2** | 0.068 | 0.5755 | 0.2499 | 0.0748 | -3.4946 | -0.0117 |
| **03 Train RMSE** | 3,948.72 | 508.6091 | 4,971.42 | 77.6849 | 4,458.79 | 366.8062 |
| **04 Test RMSE** | 2,674.94 | 1,805.22 | 4,082.54 | 4,533.88 | 5,063.97 | 2,402.54 |
| **05 Base RMSE** | 2,100.03 | 2,100.03 | 4,705.00 | 4,705.00 | 1,770.36 | 1,770.36 |

**XGBoost Regressor**

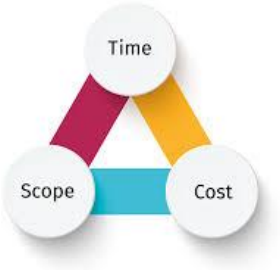
**Top features** (Linear Regression)

| **Category A** | | **Category B** | | **Category C** | |
|---|---|---|---|---|---|
| premium_s1 | 14000 | quota_ema3 | -24000 | premium_s1 | 11000 |
| quota_ema3 | -13000 | premium_s1 | 15000 | cpi_s5 | 7500 |
| bids_success_ema3 | 11000 | bids_success_ema3 | 13000 | cpi_s1 | 5900 |

**Each category with it's own model and features set, achieved RMSE <5K**

# Conclusion

Although we have achieved RMSE of < 5K. These models are far from perfect, but they still fulfil the requirement of budgetary guidance. As with all corporate projects it is always a balance of Scope, Time and Cost.




**Limitations**

- Train on limited data, have not seen more than one ten-year cycle.
- Limited features, need more features especially for Category C.
- Impact of events, i.e. Covid-19 and Financial Crisis.
- LTA reclassify the Categories

**Recommendations**

- Collecting more data (More cycles, additional features household income, unemployment rate)
- Encoding events into features (Loan restrictions, Covid-19, etc)
- Sentiment of owning a car

**Triple constraints of projects: Scope, Time and Cost**


Thank You