Introduction:

For this project, I built a machine learning model to predict the Certificate of Entitlement (COE) prices of vehicles in Singapore. A COE gives residents the right to own and use a vehicle for 10 years, and the government controls the number of COEs issued through a quota system. This scheme is designed to regulate vehicle ownership and manage road congestion in Singapore. However, COE prices fluctuate significantly due to changing demand and supply during the bidding process, making it an interesting target for prediction. My goal was to make COE prediction easier for potential vehicle owners looking to bid for a COE.

While researching the topic, I came across a paper titled “Determinants of Certificate of Entitlement Premium for Cars under Vehicle Quota System in Singapore” (Qiang MENG et al., 2015). The study examined the determinants of COE premiums, and proposed an autoregressive model to understand the COE price trend better. Inspired by this approach, I wanted to try building my own predictive model using machine learning. As a beginner, I aimed to keep it simpler, but still practical and relevant to Singapore’s unique vehicle quota system.

My dataset came from data.gov.sg, covering COE bidding results from 2010 onwards, including details such as the month, bidding exercise number, vehicle class, quota, number of bids received, and the final premium price.

Data cleaning and feature selection:

Many of the dataset’s columns were stored as strings, so the first step was to clean and standardise these columns to numeric values. This included:

- Removing commas from numeric columns such as **quota, bids_received,** and **premium.**
- Splitting the **month** column (initially in YYYY-MM format) into two separate **month_num** and **year_num** columns.
- Encoding vehicle classes (A, B, C, D, E) into numeric values (0, 1, 2, 3, 4).

The final features (X) that I selected for my model were:

- year_num
- month_num
- bidding_no
- vehicle_class

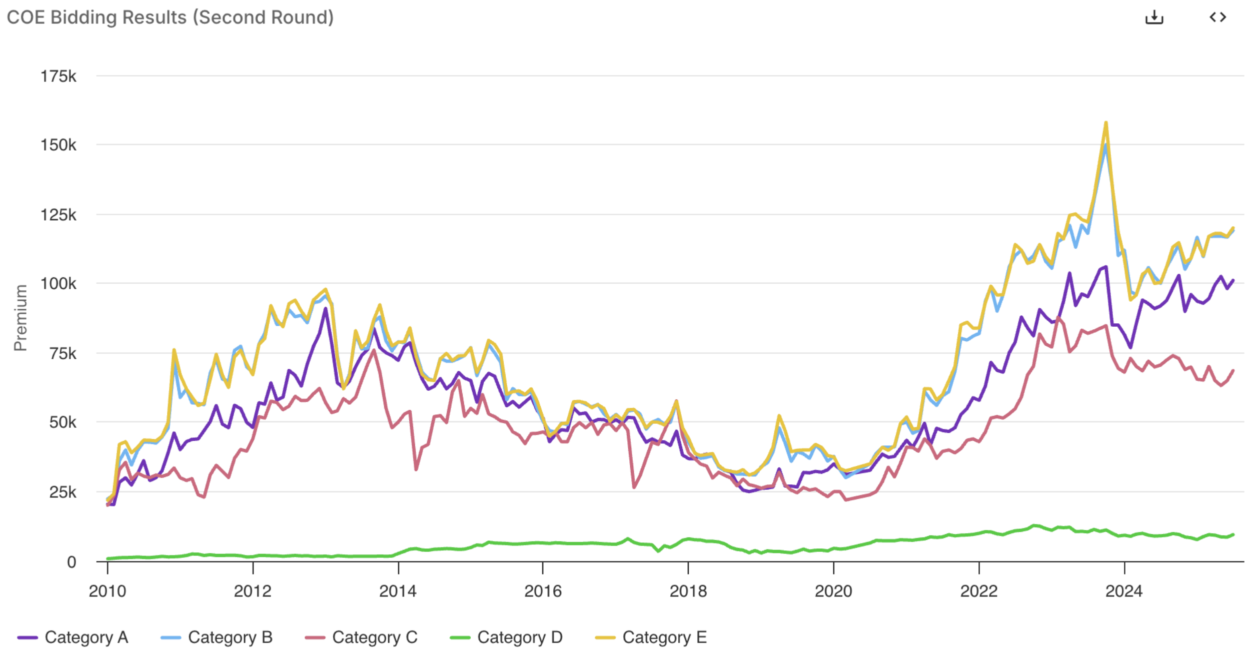
- quota
- bids_received

The target (Y) was the ‘premium’, the COE price.

The dataset was then split into training/testing data by 80/20 split.

Model selection:

Since COE prices are a single continuous variable rather than categories, I used a regression model for this project. Before selecting the type of regression model to be used, I analysed the data. The dataset source visualised the COE bidding results over the years from 2010 to present date as shown in the graph below:



Looking at the COE data over this duration, the trend is not linear, with heavy fluctuation and sharp spikes and drops. Due to the irregularity in the trend, I did not choose a linear regression model as it would not be able to capture the complex non-linear changes visualised above.

Thus, I used the Random Forest Regressor made of many decision trees which split the data into smaller regions, suitable for such non-linear trends. Random Forest also uses bagging and this reduces overfitting and gives more stable predictions compared to using a single tree.

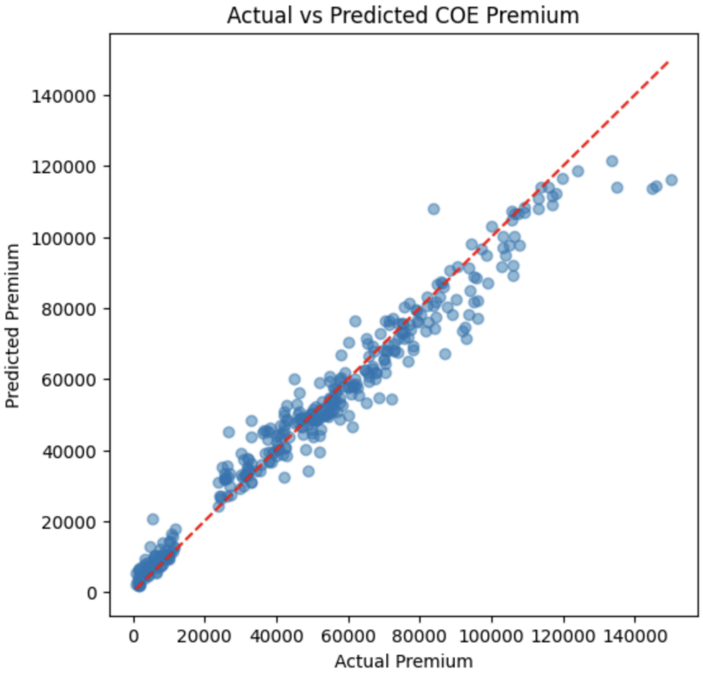
Insights:

Upon training and testing the model without tuning the Random Forest I got the following results:

- MAE: 4534.076 (3dp)
- RMSE: 6663.767 (3dp)
- R squared: 0.959 (3dp)

While R squared is very close to 1, indicating a strong prediction match, the values of MAE and RMSE were still a bit higher than expected. I believe this is because the COE prices are prone to fluctuating unpredictably due to sudden changes in demand and quota, making extreme spikes harder to predict. To try improving the accuracy, I tried using GridSearchCV to tune the Random Forest's hyperparameters, but the results stayed the same. This could have happened because the default parameters of the Random Forest were already close to optimal for this particular dataset.

Yet, overall the model performed quite well, explaining almost 96% of the variation in COE prices and generally capturing the accurate trend. There were slight deviations at very high COE values where the dataset was more sparse and volatile.



Graph derived from Final Project Colab file.

Lastly, I also wanted to understand which features from the selected few were the most useful in helping predict the COE prices. **vehicle_class** was the most significant feature, with an importance of 0.397393, followed by **year_num** (0.260905) and **quota** (0.195281). The importance of the other features is depicted in the table below.

Feature	**Importance**
vehicle_class	0.397393
year_num	0.260905
quota	0.195281
bids_received	0.120570
month_num	0.022697
bidding_no	0.003154

Conclusion:

Through this project, I learned how to source, clean and prepare real-world data, choose an appropriate machine learning model and evaluate it using regression metrics. The data revealed that COE prices can be predicted fairly accurately using features such as the quota, bids received, year, month, and vehicle class, as these features quite directly affect the COE price. I was able to achieve the goal of building a model that can predict COE prices based on past bidding patterns. My main takeaway is that while the model can capture overall trends and give good estimates based on past data, at the level I have learned ML, it still struggles slightly to predict sudden spikes due to the volatile nature of the data.

Reflection:

This course has been a great introduction to machine learning and data science for me. I learned a lot of necessary skills as a complete beginner, from cleaning up data in various ways, to the selection and evaluation of models with the correct metrics. Although it was challenging at first, it was very rewarding to reach a point where I could independently choose and build an entire ML project. Using beginner-friendly tools like Google Colab and the various python libraries, class discussions, and also the in-class assignments helped me learn and get practical experience in machine learning, greatly deepening my understanding. What inspired me most was seeing how data and models can uncover patterns that aren’t immediately obvious to the human eye, which has made me even more curious about machine learning and its possibilities. After this course, I’m excited to keep exploring AI and ML in future projects.

One thing I would advise future students to do to thrive in this course is learn some Python beforehand and also read all the extra reading materials provided in class to make the most of your learning!

Link to Colab:
https://colab.research.google.com/drive/1ShIu3iV7kxNVisXGMGR-UcGOddofYNpA?usp=sharing

References:

1. Qiang MENG, LU, Z., & Muhammad Izwan Bin OHTMAN. (2015). Determinants of Certificate of Entitlement Premium for Cars under Vehicle Quota System in Singapore. *Journal of the Eastern Asia Society for Transportation Studies*, *11*, 126–140. https://doi.org/10.11175/easts.11.126