I. BUSINESS UNDERSTANDING

In 2022, Australia was the 13th largest automotive market in the world, selling almost 1.1 million new vehicles annually (F&I Tools, 2022). If companies are able to accurately pinpoint which customers are going to make new car purchases, they will have a massive competitive advantage and this can be a game changing for the business and industry as a whole. The automotive industry plays an important role in Australia's economy. There were over 350,000 people employed within the industry and contributing 2.2 percent or over $37 billion to the nation's GDP in 2016 (Automotive Dealer Magazine, 2018).

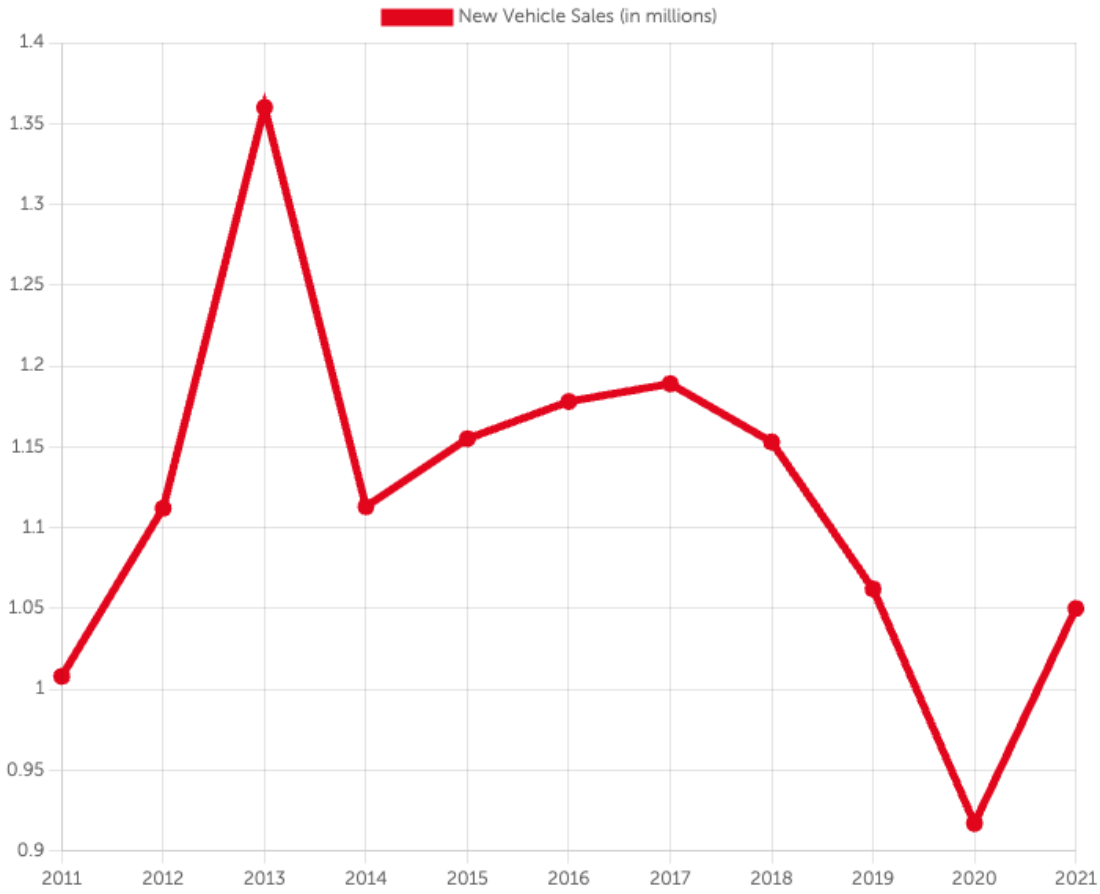

Figure 1: Sales of new vehicles in Australia (Budget Direct, 2022)

We can see in figure 1 that new car sales are bouncing back post-covid, crossing the 1 million mark in 2021. Even though countries like China and the US are still far ahead in terms of number of vehicles sold annually, but in terms of cars per capita, Australia is among one of the top countries in the world:

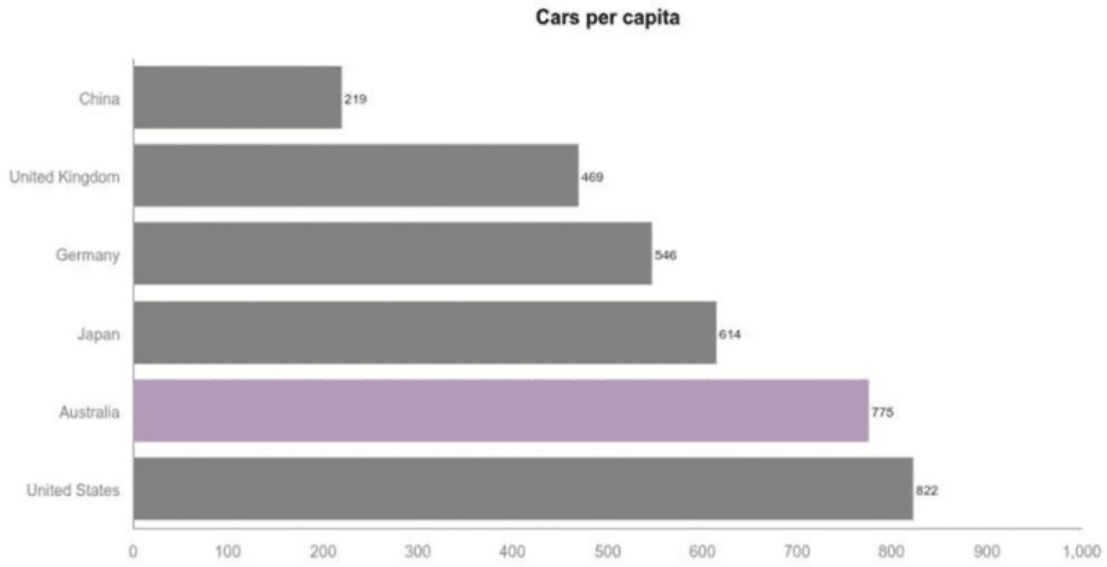

Figure 2: Cars per capita comparison among countries (Chesterton, 2022)

As of 31 January 2021, there were over 20 million registered vehicles in Australia (Australian Bureau of Statistics, 2021), with over 1 million new vehicles sold in 2021 alone. This figure is a slight increase from over 900k sold in 2020 (International Trade Administration, 2022). Type of vehicles include SUVs, commercial and passenger cars with the following breakdown:

Class	% of New Vehicle Sales in 2021	% of New Vehicle Sales in 2020
SUV	50.65%	49.59%
Light Commercial	24.12%	22.42%
Passenger	21.10%	24.22%
Heavy Commercial	4.13%	3.77%

Figure 3: Type of vehicles sold breakdown in Australia (Budget Direct, 2022)

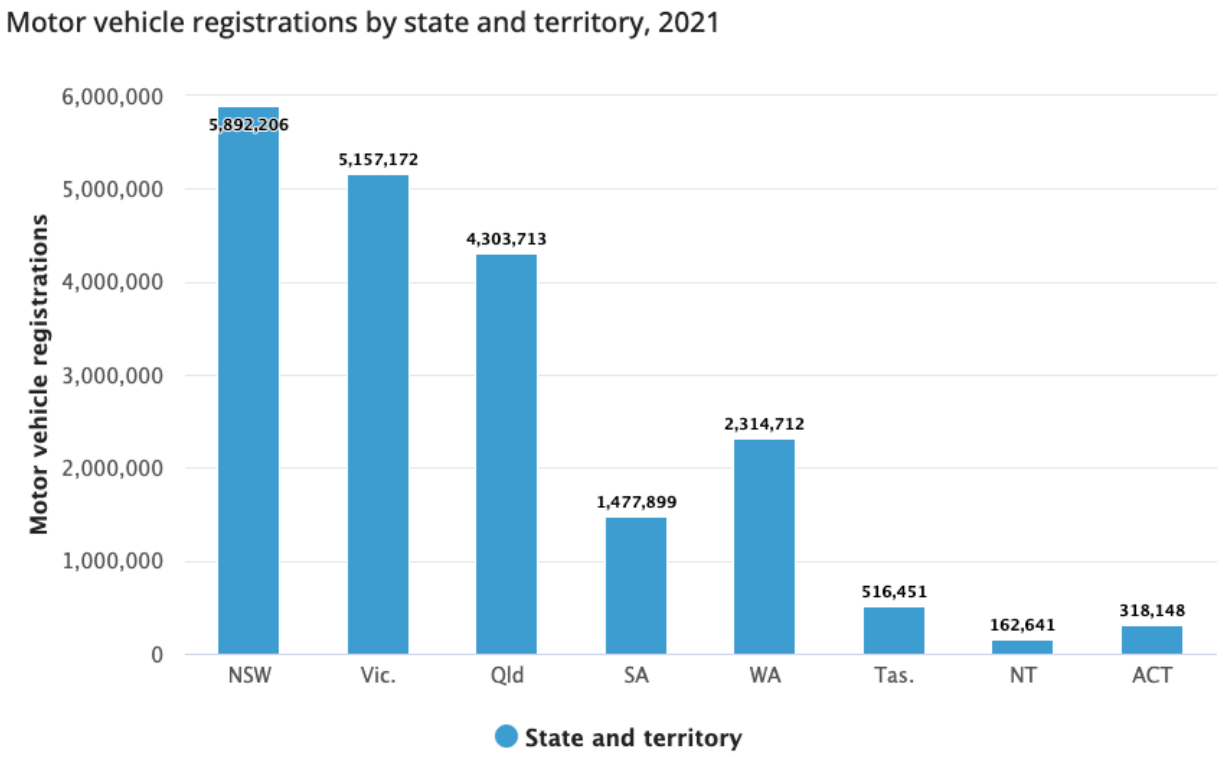

Figure 4: Total vehicles by state (Australian Bureau of Statistics, 2021)

Having a tool to predict how likely customers are going to purchase a car have several key advantages, especially if you are in New South Wales, Victoria or Queensland as shown in above figure 4, including:

1. **Sales forecasting:** Being able to predict which customers are more likely to buy a car will help the business forecast sales more accurately and will in turn, help craft the overall business strategy

2. **Targeted marketing:** Marketing team will be able to create campaigns targeting specific customers, this will improve the success chance and also saves time and resources.

3. **Inventory management:** Accurate sales forecast will help the inventory team better manage their stocks and enhance the management of their supply chain.

4. **Customer segmentation:** Knowing which customers are more or less likely to purchase a vehicle, marketing teams can segment their customers and tailor their marketing strategies to specific groups.

5. **Product development:** The idea or reasoning behind the prediction can help businesses better understand their customers needs and therefore be able to develop products that are better suited to the customers.

In this project, we will explore several machine learning models to try to predict which customers are more likely to buy a car.

II. DATA UNDERSTANDING

SUMMARY

As shown below, the features have been scaled/standardized to 1-10 in the original dataset which can be beneficial when training the model, it provides consistency and simplifies model interpretation which in turn can improve the model's performance. There is gender and age information in the dataset, we consider removing them to avoid potential bias issues. Here is the screenshot of the dataset:

	ID	Target	age_band	gender	car_model	car_segment	age_of_vehicle_years	sched_serv_warr	non_sched_serv_warr	sched_serv_paid
0	1	0	3. 35 to 44	Male	model_1	LCV	9	2	10	3
1	2	0	NaN	NaN	model_2	Small/Medium	6	10	3	10
2	3	0	NaN	Male	model_3	Large/SUV	9	10	9	10
3	5	0	NaN	NaN	model_3	Large/SUV	5	8	5	8
4	6	0	NaN	Female	model_2	Small/Medium	8	9	4	10
5	7	0	NaN	Male	model_5	Large/SUV	7	4	10	5
6	8	0	1. <25	Male	model_3	Large/SUV	8	2	8	2
7	9	0	NaN	Male	model_6	Small/Medium	7	4	9	6
8	10	0	NaN	NaN	model_4	Small/Medium	1	2	1	1
9	11	0	NaN	NaN	model_4	Small/Medium	3	1	1	2

Figure 5: Summary of the dataset

FEATURES

The dataset contains information about existing customers including car models, age of the vehicles and number of dealers visited, below is the complete list of the features:

```
 #   Column
---  ------
 0   ID
 1   Target
 2   age_band
 3   gender
 4   car_model
 5   car_segment
 6   age_of_vehicle_years
 7   sched_serv_warr
 8   non_sched_serv_warr
 9   sched_serv_paid
 10  non_sched_serv_paid
 11  total_paid_services
 12  total_services
 13  mth_since_last_serv
 14  annualised_mileage
 15  num_dealers_visited
 16  num_serv_dealer_purchased
```

Figure 6: list of columns in the dataset

There are 17 columns (including target variable) and 131,337 rows in the dataset.

DATA FORMAT

Majority of the columns are integers and some are categorical as shown in Figure 7:

```
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   ID                       131337 non-null  int64
 1   Target                   131337 non-null  int64
 2   age_band                 18962 non-null   object
 3   gender                   62029 non-null   object
 4   car_model                131337 non-null  object
 5   car_segment              131337 non-null  object
 6   age_of_vehicle_years     131337 non-null  int64
 7   sched_serv_warr          131337 non-null  int64
 8   non_sched_serv_warr      131337 non-null  int64
 9   sched_serv_paid          131337 non-null  int64
 10  non_sched_serv_paid      131337 non-null  int64
 11  total_paid_services      131337 non-null  int64
 12  total_services           131337 non-null  int64
 13  mth_since_last_serv      131337 non-null  int64
 14  annualised_mileage       131337 non-null  int64
 15  num_dealers_visited      131337 non-null  int64
 16  num_serv_dealer_purchased 131337 non-null  int64
```

Figure 7: Additional information about the columns

MISSING VALUES

As shown in figure 7, there are many missing values in ‘age_band’ and ‘gender’ columns; almost half of the values in each column were missing.

OUTLIERS

We didn’t detect any outliers in the dataset since all the features have been scaled to 1-10. Example below:

```
df['sched_serv_warr'].value_counts()
```

```
2     13788
1     13484
3     13305
4     13129
9     13119
6     12972
8     12938
10    12907
5     12859
7     12836
```

DUPLICATE VALUES

No duplicate values were removed considering the majority of the features are numerical values and due to the nature of the features, all of them (example: car models, car segments, months since last serviced) are expected to have some identical values.

III. DATA PREPARATION

MISSING VALUES

‘Age_band’ and ‘gender’ columns were removed due to many missing values (more than half of the values were missing) and also to avoid potential bias issues:

```
df_cleaned.isna().sum()
```

```
Target                      0
car_model                   0
car_segment                 0
age_of_vehicle_years        0
sched_serv_warr             0
non_sched_serv_warr         0
sched_serv_paid             0
non_sched_serv_paid         0
total_paid_services         0
total_services              0
mth_since_last_serv         0
annualised_mileage          0
num_dealers_visited         0
num_serv_dealer_purchased   0
```

CATEGORICAL COLUMNS

'Car_model' and 'car_segment' contained categorical values, we encoded the values, effectively converting them into numerical values so we can use them to train models.

Car_model:

```
df_cleaned['car_model'].unique()
```

```
array([ 0.,  1.,  2.,  3.,  4.,  5.,  6.,  7.,  8.,  9., 10., 11., 12.,
       13., 14., 15., 16., 17., 18.])
```

Car_segment:

```
df_cleaned['car_segment'].unique()
```

```
array([0, 1, 2, 3])
```

DATA SPLITTING

We split the data into 3 smaller sets for training, validation and testing purposes, as follows:

-Training set (60%)
-Validation set (20%)
-Test set (20%)

IV. MODELING

We trained and tested 6 different algorithms and experiments to see which model fits better, in each experiment, we tested several different hyperparameter tunings aiming for optimal accuracy:

1. **Experiment 1:**
   Logistic Regression: a popular statistical technique that is relatively simple, interpretable, flexible and computationally efficient
2. **Experiment 2:**
   KNN: flexible algorithm and useful in cases where data is non-linear and/or has more complex feature relationships, although it has a higher computation cost
3. **Experiment 3:**
   SVC: powerful algorithm for classification problems especially with dataset containing noisy data or complex feature relationships
4. **Experiment 4:**
   Decision Tree: useful and flexible algorithm for machine learning applications, can handle highly non-linear data or has complex feature relationships
5. **Experiment 5:**
   Random Forest: a collection of decision trees trained on different subsets of data and features. It aggregates the prediction outcomes of all individual decision trees for its final prediction, which can be computationally expensive.
6. **Experiment 6:**
   ExtraTrees: a variation of Random Forest aiming to reduce the model variance by increasing the randomness of individual trees, the downside is that the randomness can lead to difficulty in interpreting the model.

ASSESSMENT METHOD:

1. **Prediction score:** >80%
2. **Recall score:** >80%
3. **F1 score:** >80%
4. **Low overfitting**

-Ideal scenario: the model(s) performs exceptionally well in all three metrics and does not overfits, which would be advantageous for the business

-Intermediate scenario: The model(s) achieves a high score in one or two of the metrics, which can still benefits the business, for example:

- **High Precision score:** The model prioritizes identifying customers who are likely to buy a new car, but may miss some actual buyers. This allows us to allocate marketing and sales resources effectively, targeting actual buyers and saving company resources. However, this could also mean missing out on some potential buyers.

- **High Recall score:** The model prioritizes identifying all new car buyers, even if it results in a higher number of false positives. This approach helps us identify more buyers and engage with them, generating more revenue in the long run. However, this could also result in resources being spent on non-buyers.

-Worst-case scenario: The model(s) performs poorly in all three metrics, indicating that it is not suitable for our business's needs. In such cases, we will need to explore alternative algorithms that offer better prediction accuracy.

V. EVALUATION

We performed many hyperparameters tuning on each of the algorithms with varying results as shown in the individual reports (please see experiment reports for details), some models performed better based on precision score, others had better recall scores.

Ideally, every business would want a model that achieves the highest f1 score, which would mean the model is also performing well both in precision and recall. With that assumption, we have summarized the best performing model for each algorithm as shown below:

	LogReg	KNN	SVC	DecisionTree	RandomForest	ExtraTrees
Precision						
Training	0.82842	0.95144	0.91496	0.92363	**0.97548**	0.99805
Validation	0.81752	0.90881	0.88483	0.82515	**0.92308**	0.93822
Testing	0.82558	0.91686	0.93096	0.86557	**0.94406**	0.94794
Recall						
Training	0.21207	0.6433	0.61579	0.84783	**0.82964**	0.90994
Validation	0.19893	0.53108	0.5595	0.746	**0.746**	0.72824
Testing	0.2017	0.5483	0.59375	0.75	**0.76705**	0.75
F1						
Training	0.33769	0.7676	0.73614	0.88411	**0.89667**	0.95196
Validation	0.32	0.6704	0.68553	0.78358	**0.82515**	0.82
Testing	0.3242	0.68622	0.72507	0.80365	**0.84639**	0.83743

Random Forest model outperformed other algorithms followed by Extra Trees, it also has relatively low overfitting which makes it the best model in this case. But the model's recall score is lower than target (80%), if the business is looking to predict as many potential buyers as possible, further hyperparameters tuning can help boost recall score but might be at the expense of precision score and ultimately, f1 score.

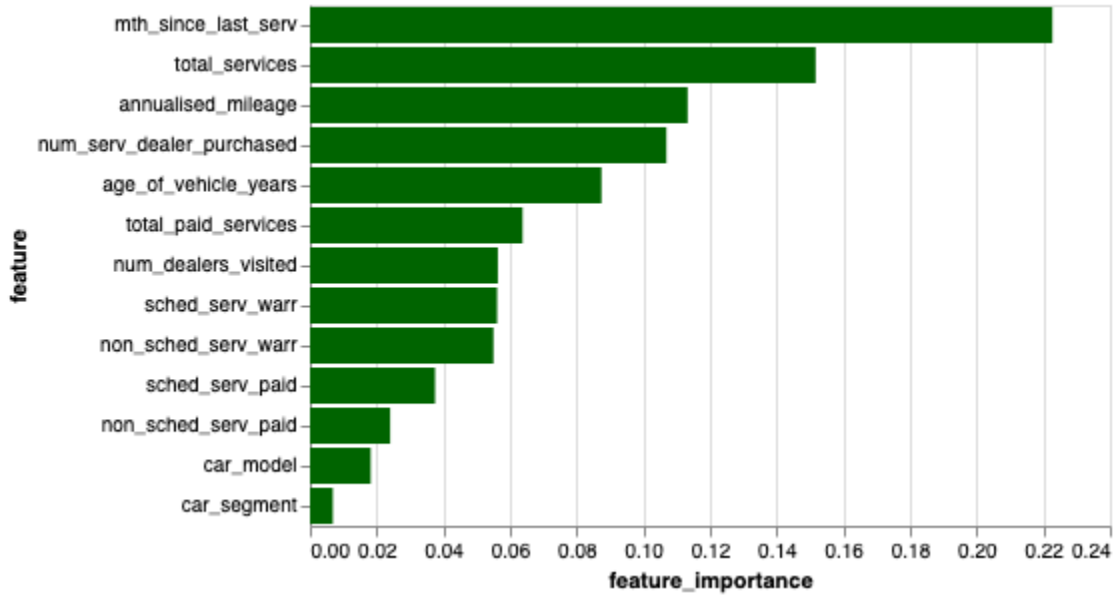

Figure 8: Features importance

We can also gain more insights by looking into the importance of each feature. Figure 8 shows the contribution of each feature to the model's accuracy in making predictions. By understanding the relationship between input features and model's output, we can identify irrelevant features which can then be removed to simplify and improve the model's efficiency.

Overall, if the business is looking for a model that performs well in both precision and recall score, Random Forest model is their best bet in this case.

VI. DEPLOYMENT

The model has a decent 84% f1 score when tested using a testing dataset. Even though the result is convincing, it would be a good idea to trial run it in a real world setting first.

By creating a thorough and concrete plan to closely monitor the model's performance, assess potential impact on the audience (example: bias), record any anomalies, gather feedback from

users and fine tune the model during the trial run will ensure the model's readiness before deploying it to a larger audience.

REFERENCES

1. Australian Bureau of Statistics. (2021, June). *Motor Vehicle Census, Australia.* https://www.abs.gov.au/statistics/industry/tourism-and-transport/motor-vehicle-census-australia/latest-release
2. International Trade Administration. (2022, December). *Australian Automotive Industry.* https://www.trade.gov/market-intelligence/australian-automotive-industry
3. F&I Tools. (2022). *Car Sales by Country.* https://www.factorywarrantylist.com/car-sales-by-country.html
4. Automotive Dealer Magazine. (2018). *INDUSTRY REPORT FEATURE: THE AUSTRALIAN AUTOMOTIVE INDUSTRY'S ECONOMIC CONTRIBUTION.*

https://automotivedealer.com.au/industry-report-feature-the-australian-automotive-industrys-economic-contribution/

5. Chesterton, A. (2022). Australian Car Market: Car Sales, Statistics and Figures. https://www.carsguide.com.au/car-advice/australian-car-market-car-sales-statistics-and-figures-70982
6. Budget Direct. (2022). *Latest Australian Car Sales Statistics & Survey.* https://www.budgetdirect.com.au/car-insurance/research/australian-car-sales-statistics.html