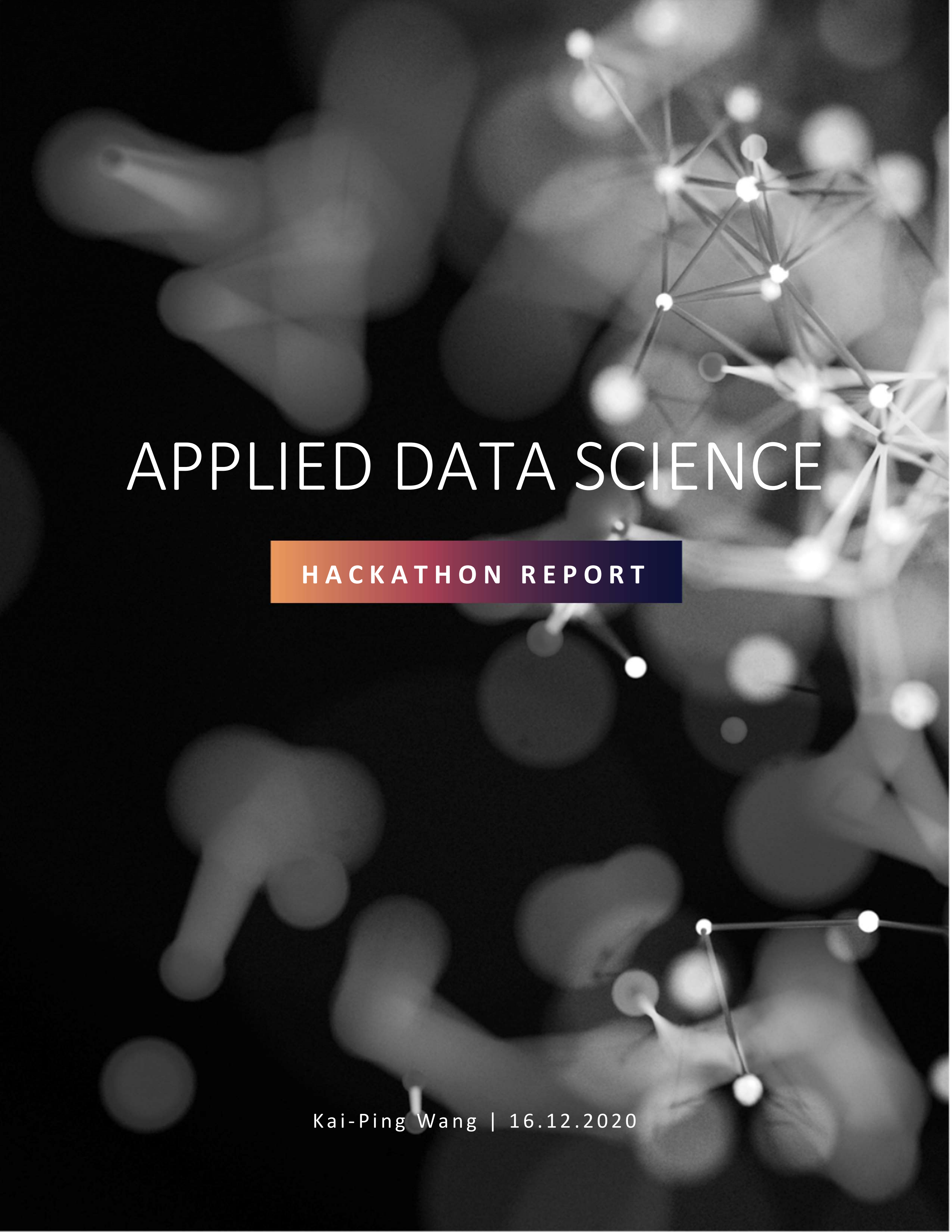
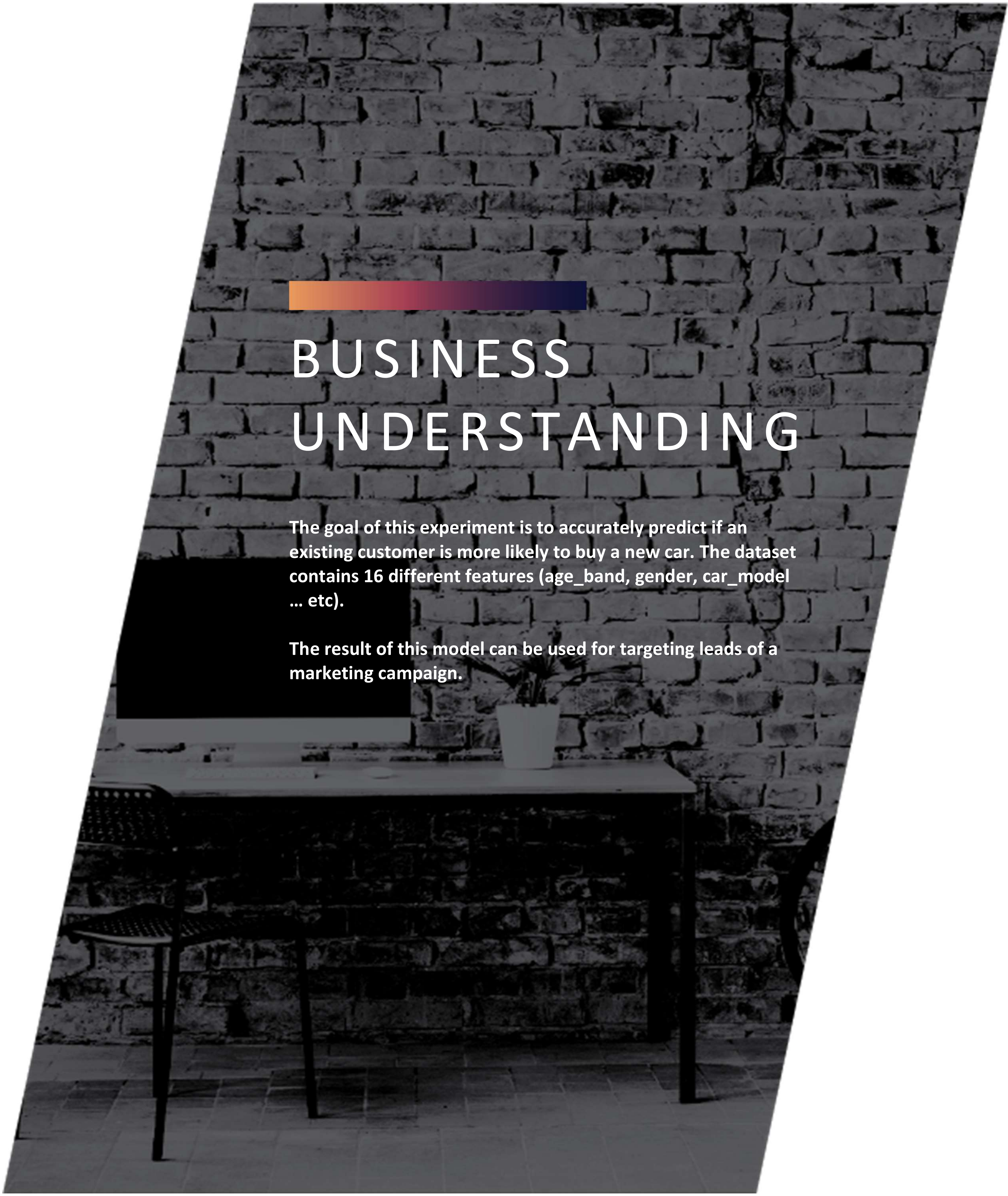


APPLIED DATA SCIENCE

HACKATHON REPORT

Kai-Ping Wang | 16.12.2020



BUSINESS UNDERSTANDING

The goal of this experiment is to accurately predict if an existing customer is more likely to buy a new car. The dataset contains 16 different features (age_band, gender, car_model … etc).

The result of this model can be used for targeting leads of a marketing campaign.

DATA QUALITY

- Only two features (age_band, gender) have missing value.
- **Target value is imbalanced** with 98:2 ratio on negative and positive result.

DATA UNDERSTANDING

COLLECT DATE

The data used in this experiment is coming from https://raw.githubusercontent.com/aso-uts/applied_ds/master/assignment2/repurchase_training.csv

DESCRIBE DATA

There are 16 features and 131,337 observations in this dataset. Most features are in numeric format, and age_band, gender, car_model and car_segment features are in string format.

```
 #   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
```

EXPLORE DATA

As the question being asked here is if a customer is likely to buy a new car, so the answer is either true or false. When checking the Target value on value counts, we can see the dataset is extremely imbalanced with 98:2 ratio. This means the performance on positive may be affected. The metric to be used for measuring the classifier performance needs to have weighted concept (F1 or MCC).

FORM HYPOTHESES

- **Customers with older cars are more likely to buy.**

  When a car being driven for a long time, there are more factors to encourage the owner to buy a new car. For example, the maintenance cost, fuel efficiency, missing out new features…etc

  Therefore, the assumption here is higher the **age_of_vehicle_years**, more likely a customer is looking to buy a new car.

- **Customers use the car frequently are more likely to buy.**

  When customers use their cars more, it indicates that cars are important in their day-to-day life, and their cars have more wear and tear. This implies that they are more likely to invest in buying a better car.

  Feature **annualised_mileage** is one indication in this area to see how much owners use their cars.

DATA PREPARATION

SELECT DATA

- Include all features.

CLEAN DATA

- Some missing data in age_band and gender. Due to the potential feature significance, instead of replacing missing value with existing value, we are using ‘OTHERS’ for missing value. By doing this, we don’t contaminate the importance of existing value.

CONSTRUCT DATA

- Label encoding all categorical data.

FORMAT DATA

- None as this is classification question.

Missing Data

- Replace missing data with ‘OTHERS’ to avoid contaminating the importance of existing value.

MODELING

TEST DESIGN

- Split training set and test set by 80:20.

MODEL OPTIONS

Select a number of classifier models as candidates, and pick the final champion in evaluation stage based on performance. Fine tune these models further using hyperparameters.

- **LogisticRegression**
- **KNeighborsClassifier**
- **RandomForestClassifier**

HYPERPARAMETERS

- **RandomForestClassifier**
    - random_state
    - criterion
    - n_estimators
    - max_depth

EVALUATION

F1 Score is used to evaluate the performance of each model included in this experiment. The goal is to find a good F1 score across both training set and test set, and not overfitting nor underfitting. **Below is the result of all tried models, and the highlighted ones are the best of each algorithm.**

	F1 - train	F1 - test
KNeighborsClassifier (default)	**0.7614**	**0.6920**
LogisticRegression (default)	0.3389	0.3392
RandomForestClassifier (default)	**0.9998**	**0.8779**
RandomForestClassifier (n_estimators=150, random_state=8, criterion='entropy', max_depth=13, min_sample_leaf=5)	**0.8748**	**0.8467**

The final pick is RandomForestClassifier with hyperparapeter tuned. This is because the difference in F1 score between training set and test set is much closer, and not too low in general. This means the model is neither overfitting nor underfitting.

Confusion Metrix

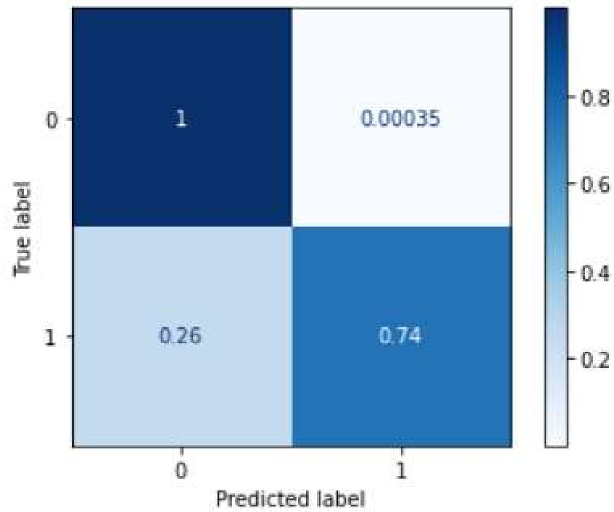
The confusion metrix on the right is derived from the final model picked from above with testing set data.

We can see that the negative prediction is very accurate, but the positive prediction is not very good in comparison.

In marketing perspective, it is not bad to include false positive into marketing campaign as they are potential prospects.

However, being able to rule out true negative with high Confidence is important for the business to save cost.

```
[[25576     9]
 [  175   508]]
```



Feature Importance

As we use Random Forest Classifer algorithm, we can extract feature importance from the model, and it is shown in the picture on the right.

How many months since last service has much higher significance compared to other features.

Gender and age_band have lots of missing value, so it is not surprising to see they have low significance in the final prediction model.

Our original assumption on age_of_vehicle_years and annualized_mileage are among the top 5 features.

	feature	importance
11	mth_since_last_serv	0.167563
5	sched_serv_warr	0.122501
12	annualised_mileage	0.117310
10	total_services	0.106073
4	age_of_vehicle_years	0.096858
14	num_serv_dealer_purchased	0.094212
7	sched_serv_paid	0.091895
13	num_dealers_visited	0.049941
9	total_paid_services	0.042831
6	non_sched_serv_warr	0.039714
1	gender	0.027532
8	non_sched_serv_paid	0.025480
2	car_model	0.011599
3	car_segment	0.004494
0	age_band	0.001997

WHAT'S NEXT

The best model based on F1 score in general is Random Forest Classifier. Although the default setting can achieve higher F1 score in general, the difference between training set and testing set is too large, which indicates the model is overfitting, and it may not be reliable for unseen data. The final model picked with hyperparameters tuned on max_depth and min_sample_leaf has much closer F1 score between training set and test set, and the performance isn't underfitting.

Looking back to the original question of this experiment, business wants a model that can predict if a customer is likely to buy a car. Due to the imbalanced dataset, the final model can achieve high accuracy on predicting negative (not buying) and not very good accuracy on positive (buying). However, this model can certainly be used for targeting leads to marketing campaign as it can rule out the true negative with high confidence.

Based on the business requirement, this model is able to rule out true negative with high confidence. Although the false positive is bit high, in marketing campaign perspective, the impact is not bad. Therefore, it is recommended to deploy this model into Production.

It is also recommended to capture more positive observation for future training to achieve higher accuracy on true positive.