36106 Machine Learning Algorithms and Applications-Autumn 2023

Assessment Task 2

Classification Models

Student Name: Shalimar Chalhoub

Student ID: 14071892

UTS

UNIVERSITY OF TECHNOLOGY, SYDNEY

Contents

Final Report for Car repurchasing Classification Models

Introduction

The aim of this project is to build a classification model that can predict which customers are most likely to repurchase a car from the company given a set of features provided such as gender, age band, the car model and segment as well as many other features . The assignment consisted of 5 parts, each testing a different hypothesis and a new way to tackle the problem in order to find the ultimate one. All of the experiments were done in Colab, using Python for data cleaning and exploration, visualizations, data analysis and building the models.

This report presents the results of the project according to the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology.

Business Understanding

Customer retention is a very important factor in increasing sales and bettering a company's revenue thus why our company wants to improve customer retention by identifying customers who are most likely to repurchase a second car and targeting them with special marketing campaigns as well as one on one marketing such as follow up calls and personalized services. We believe that by identifying which customers are most likely to repurchase, we can improve customer loyalty and generate better sales.

A successful model would give a recall score above 75% as well as a good precision score, as we do want to predict most true values as true but we also don't want to have a lot of false values predicted as true because that will lead to a loss of company resources and a lot of unnecessary costs.

By using python to preprocess our data and create our binary classification models, we are able to engineer new features, manipulate the data and create a model that can accurately predict clients that will repurchase a second car thus increasing company revenue.

Data Understanding

The data used in this project was provided in the initial phase of this project and it consist of 17 different features including information about the gender of the client, their age group as well as information about the car model and segment as well as the age of their older car, how many month since their last services as well as a lot of other features that can be seen in Appendix A.

The target variable is called 'Target' in the model and it is the customers that have repurchased a second car from the company, with 0 representing that they haven't and 1 representing that they have.

It is important to note that our data is extremely imbalanced with 127816 customers that have not repurchased a car randomly 3521 which is even less than 1%.

Below in figure 1, we can see the repartition of age_band values when the target is 1, below is the dictionary of age_band

"1. <25": 1,
"2. 25 to 34": 2,
"3. 35 to 44": 3,
"4. 45 to 54": 4,
"5. 55 to 64": 5,
"6. 65 to 74": 6,
"7. 75+":7

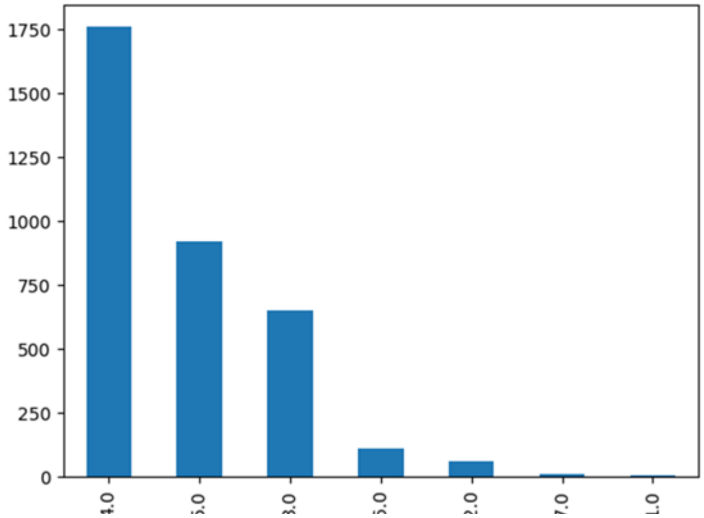

Figure 1- Graph representation of age_band when Target is 1

From figure 1, we can see that the age_band most likely to repurchase is age_band 4, which are individuals between the age of 45 and 54, followed by band 5 which are individuals between the age of 55 and 64. Theoretically, this data is accurate as these are the ages where typically, individuals would repurchase cars

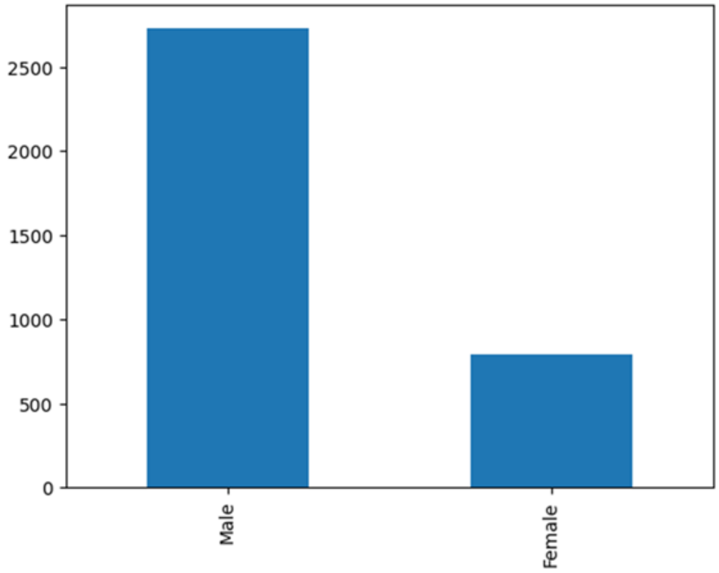

Figure 2- graph representation of gender when Target is 1

Figure 2 shows the distribution of gender for individuals that have repurchased cars and It can be seen that male customers who repurchases are almost 3 times as much as females

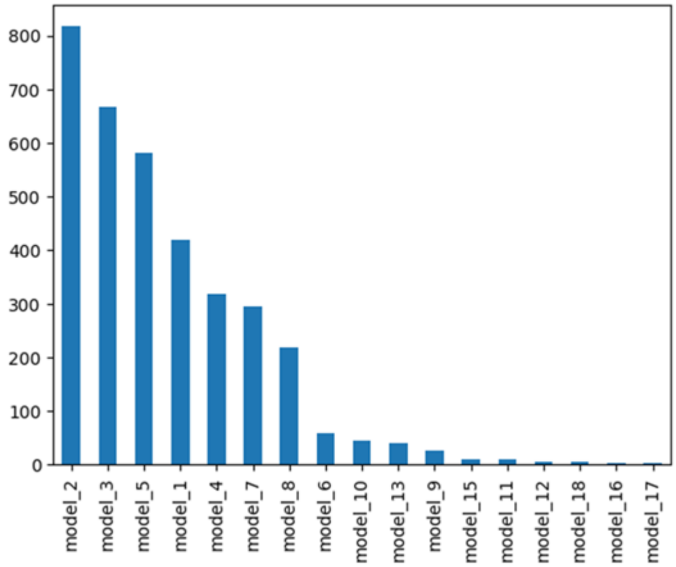

Figure 3- graph representation of car model when Target

Figure 3 shows that the car models purchased the most are models 2, 3, 5, 1 and 4 and the ones purchase the least are models 17, 16, 18 12, 11 and 15. It can also be seen that there are a lot of fluctuations between the values

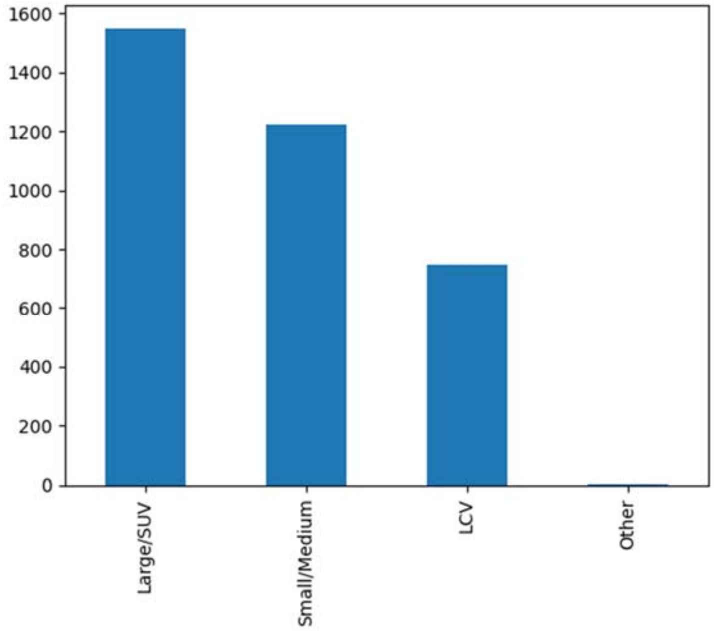

Figure 4- graph representation of car segment when Target

From figure 4, we can see that the car segment repurchased most is a large car or SUV and the least is LCV or other.

Ethics and Privacy

It is very important to look at ethics and privacy when dealing with data such as age and gender. It is essential to be cautious and not force individuals to reveal their age or gender, while also avoiding making assumptions about anyone's gender as these are ethical matters that should be taken into consideration and are of utmost importance.

Furthermore, it is crucial to recognize that gender is not limited to just male and female. We should aim to incorporate all genders or at least include an "other" column to ensure that everyone feels represented and included or we can even add “prefer not to say”.

To safeguard privacy, such personal data must be kept confidential and even encrypted or disguised to avoid any risk of re-identification. Protecting privacy should be a top priority when dealing with sensitive personal data.

Data Preparation

All the experiments had the same data preparations done on them which consisted of the following:

1. Dropping the ‘ID’ column from the dataset as it causes overfitting
2. Checking for duplicated rows and dropping them from the dataset as well
3. Checking if any car models had more than 1 segment. I found that model_17 had both ‘other’ as well as ‘small/medium’ car segments.
4. Changing the ‘other’ car segments of model_17 to ‘small/medium’ and updating it in the dataset
5. Resetting the index of the data, just to have better view of the data
6. Extracting the column ‘Target’ to a different dataframe called y
7. Mapping age_band and gender to convert them to numerical
8. Dividing categorical columns and numerical columns
9. Performing one hot encoding on the numerical data
10. Concatenating the numerical and categorical data into one dataset
11. Splitting into train dev and test sets
12. Using a KNN Imputer to impute missing data from each of the sets
13. Rounding the missing data to the nearest integer

Modeling

For Part1, I first balanced the data using SMOTE with a sampling strategy of 0.7 and then did a LogisticRegression with the following hyperparameters:

- max_iter = 800,
- random_state =42
- C=0.5

For Part2, I sampled the data using EditedNearestNeighbours and then applied a KNeighborsClassifier with the following parameters:

- n_neighbors = 22
- metric = ‘euclidean’

For Part3, I decided to train a DecisionTreeClassifier with the following hyperparameters

- min_samples_split=30
- max_depth=15

For Part 4 I did a RandomForestClassifier with the following hyperparameters:

- n_estimators =100
- max_depth=15
- min_samples_leaf=10
- max_features = None

For Part 5, I also did a random forest but I combined it with GridSearchCV and it gave me the following best parameters

- n_estimators =30
- max_depth =25
- min_sample_leaf=10

Evaluation

Below are the results for the test set of each model, we will be looking at the recalls scores:

Part 1: 0.80 (however, for this model it is unacceptable as it is paired with a very bad precision score)

Part 2: 0.54

Part 3: 0.75

Part 4: 0.73

Part 5: 0.59

The model that is performing best is obviously model 3 which is the decision tree and it can be seen from the figure below as well

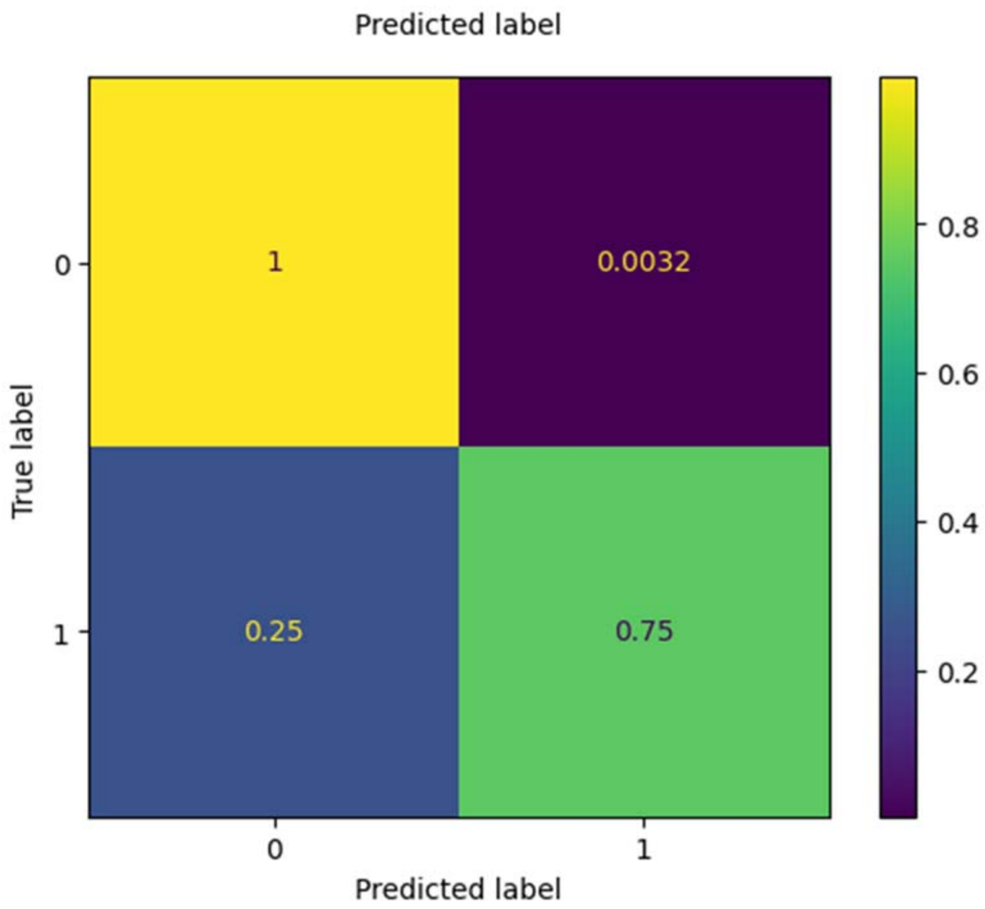

Figure 5- Decision Tree Matrix

It can be seen above that the True Negatives show as 1 however, it is important to note that this is not the case, the data is just severely imbalanced. This number is supposed to be 0.9968

It can be seen that this model works the best with data a True positive rate of 75% which would work well for the business and generate good income without having too much loss on inaccurate data.

Deployment

This model will be deployed and developed as Python code. It will particularly benefit the marketing and retails field whose stakeholders include car sales companies, insurance companies as well as other businesses to leverage the model's predictive capabilities and identify individuals who are likely to repurchase a second car based on specific features from their first purchase. This model is also useful when working with skewed or imbalanced, however, not linear data.

By identifying individuals who are likely to repurchase a second car, companies can tailor their marketing strategies to these individuals, potentially increasing the likelihood of future sales and overall profits. The implementation of this model has the potential to bring significant value to companies in the automotive industry and beyond, enabling them to make data-driven decisions and optimize their operations.

References

For all of the models I have used the templates provided in the MLAA sessions

Appendix

Data Dictionary:

ID: Unique ID of the customer

target: Model target. 1 if the customer has purchased more than 1 vehicle, 0 if they have only purchased 1.

age_band: Age banded into categories

gender: Male, Female or Missing

car_model: The model of vehicle, 18 models in total

car_segment: The type of vehicle

age_of_vehicle_years: Age of their last vehicle, in deciles

sched_serv_warr: Number of scheduled services (e.g. regular check-ups) used under warranty, in deciles

non_sched_serv_warr: Number of non-scheduled services (e.g. something broke out of the service cycle) used under warranty, in deciles

sched_serv_paid: Amount paid for scheduled services, in deciles

non_sched_serv_paid: Amount paid for non scheduled services, in deciles

total_paid_services: Amount paid in total for services, in deciles

total_services: Total number of services, in deciles

mth_since_last_serv: The number of months since the last service, in deciles

annualised_mileage: Annualised vehicle mileage, in deciles

num_dealers_visited: Number of different dealers visited for servicing, in deciles

num_serv_dealer_purchased: Number of services had at the same dealer where the vehicle was purchased, in deciles