Group Assignment 3

The Due date for this assignment is Feb 27th

Write the name of all your group members at the top of the report.

Attach all pieces of the code to the respective part of the question.

Whenever you run a logistic regression or a similar analysis, take a screenshot of the R output and report it in the respective part of the question.

In this assignment, you are tasked with testing predictive models to maximize a campaign profit for the company. The focal company is involved in the business of providing soda machines to different eateries such as fast food and restaurants. The company has already implemented a campaign by sending an offer regarding the new version of the machine to their customer base. However, analysis of response from the campaign is not very promising as many food establishments have decided to stick with the old version of the machine instead of getting a new version. This is despite the fact that the company offers some discount on the new machines.

The marketing team analyzed the results of the first campaign and decided to run a "follow-up" campaign in order to give customers who said no in the first campaign another chance to get the newer version of the product. However, the challenge is that the potential target audience already turned down the company's offer and hence it is less likely that a second attempt to be successful. From historical data, the company believes that we should expect a **follow-up campaign response rate of only 50% of the response rate of the first campaign.**

At the current cost and profit structure, targeting everyone might not be the most profitable move. Therefore, the company has decided to choose a very conservative strategy: instead of running the follow-up campaign test in a randomized way which can be costly, the marketing team plans to **infer** the profitability of the follow-up campaign **solely** based on the results of the first campaign. The cost of using a new salesperson to call the customer and explain the benefit of the product is estimated to be $20 per customer, and if the customer says Yes to the follow-up campaign, the company can expect to earn a profit of $786 for the lifetime of the customer.

Your job as a data analyst in the marketing team is to decide who to target for the follow-up campaign. Your final objective is to maximize profit of the follow-up campaign. The dataset contain information about the food establishment and the response column is "buy1" column which shows if the customer has bought the product in the first campaign or not.

Using this variable as your dependent variable, run predictive models where they are trained on the training data and tested on the test data. The data has a column called "training" where value 1 means that the customer is part of the training data and 0 means test data. As a result, do NOT create your own training and test data based on randomization, instead use the training column to see who should be in training and who should be in test. **Your profit calculations must all be done on the test dataset ONLY**. In other words, once you train your model on the training data, everything else should be calculated on the 15,000 customers in the test data only.

Question 1 – use logistic regression, decision trees, pruning, bagging, random forest, and neural networks to decide which customers in the test data should be targeted. For each model, calculate the follow-up campaign profit and compare all the models. You are free to use any variables or functional forms or neural network structure that you believe can help your predictions. **Your grade will be proportional to the reported highest follow-up campaign profit of your models. The higher the profit, the better score you can achieve on this assignment.**

Couple of important notes:

1) Here we do not do classification in the sense of getting 0 and 1 labels based on a threshold such as 0.5 or any other threshold. Instead, we use prediction probabilities and compare them to break-even response of the follow-up campaign to decide who should be targeted. For the decision tree, you can use the following command to get probabilities
   `test_data$predictions <- predict(tree_model, newdata = test_data, type = "prob")`
   and for the case of bagging/random forest you can use a command such as:
   `bagging_preds <- predict(bagging_model, newdata = test_data, type = "prob")`

2) This assignment has only one question and 14 points. As a result, do not be shy about writing about all the models you have tested with a complete picture of any initial model you tested out and any adjustments you made to it in order to increase its profitability predictions. By writing a complete report on your efforts, you increase the likelihood of partial credit if parts of your calculations are not correct.
3) Do NOT use zipcode in the data as a categorical variable. There will be more than 20,000 zipcodes in the data and given the size of the data and the free resources we

have at our disposal (Google Colab free version) we would be unable to use 20,000 fixed effects.

Data Description

cust_id	Customer id
state	US state where the customer is located
zip	5 digit ZIP code
speeddown	Average download broadband internet speed in the ZIP code
speedup	Average upload broadband internet speed in the ZIP code
last	Time (in months) since the customer's last order
numords	Number of orders in the last year
dollars	Total money spent (in hundreds of dollars) in the last 4 years
sincepurch	Time (in months) since most recent purchase
refurb	Is 1 if the customer has ever purchased a refurbished machine
oldmodel	Is 1 if the customer's most recent machine purchase is a model no longer sold

eightvalve	Is 1 if the customer's most recent machine purchase has 8 different nozzles
type	Type of food establishment
income	Median household income (100k dollars) in ZIP code
medhvalue	Median value (100k dollars) for all owner-occupied housing units in ZIP code
buy1	Response to wave 1 offer (1 = accepted the offer, 0 = did not accept)
training	70/30 split (1 = training sample, 0 = validation sample)

Data pre-processing code for neural network in Google Colab:

Once you open your data into Colab, you can run the following code in Colab to pre-process the data before getting ready to run neural networks:

```
features = data.drop(columns=['cust_id', 'buy1', 'zip'])

# One-hot encode categorical variables
features = pd.get_dummies(features, drop_first=True)  # Convert categorical variables

columns_to_scale = ["last", "numords", "dollars",
```

```
                "sincepurch", "income", "medhvalue"]
scaler = StandardScaler()
features[columns_to_scale] = scaler.fit_transform(features[columns_to_scale])


data_processed = pd.concat([features, data['buy1']], axis=1)


trainData = data_processed[data_processed['training'] == 1]
testData = data_processed[data_processed['training'] == 0]


trainData1 = trainData.drop(columns=['training','buy1']).to_numpy()
testData1 = testData.drop(columns=['training', 'buy1']).to_numpy()
train_labels = trainData['buy1'].to_numpy()
test_labels = testData['buy1'].to_numpy()
trainData = trainData1
testData = testData1


trainData = trainData.astype(np.float32)
testData = testData.astype(np.float32)
```

Now the data is ready for neural networks. Train the network on trainData and get predictions on testData. Add predictions as a new column to your testData:

testData['predicted_prob'] = pred_probs

where pred_probs here stores the predictions from neural network. You can then export the testData as a csv and work with it in python:

from google.colab import files

testData.to_csv('nn.csv', index=False)

Download the file

files.download('nn.csv')

you can now import nn.csv file from your Downloads folder into R and continue with economic analysis.

Good luck!