Data Project

# Defining the question

The problem which I am trying to solve can be defined as: “Build a model using logistic regression, in order to predict the probabilities of different segments of customers responding to different campaigns, and indeed discover what these segments are. Furthermore, I will look at which campaign/s were the most successful, in order to answer the question of which campaigns should be used in the future and who should they be targeted to.”

I decided to use a logistic regression algorithm since it is not too prone to over-fitting. It can give an understanding not only of the relevance of a variable in predicting the outcome, but also the positive or negative affliction of this relevance. It is a relatively easy method to implement and train.

# Data Cleaning and Preparation

Next, to approach this task, I went about cleaning and preparing the data. I made the assumption that every person in this dataset had been exposed to every marketing campaign, and that a 1 in the “Campaign\_n” column meant that they had responded positively to that campaign, and a 0 that they had not responded; although for a real project this would need to be verified. I also decided to delete any data which I wasn’t going to use in the analysis: I deleted most fields apart from “Subscriber ID”, “Campaign\_n”, “Customer Type Group”, “Customer Segment”, “Sales Channel”, “Gender”, “Birth date”, “Marital Status”, “Nationality”, and “Occupation”, as I couldn’t confidently discern their meaning or relevance. I did this deletion in MS Excel as I was already browsing the data in Excel.

I decided to segment the Birth Date values into the following segments: 18-25, 26-35, 36-45, 46-55, 56-65, and 66+ as these are the most commonly used age segments for marketing data.

# Data Exploration

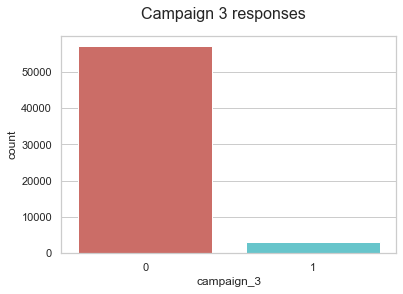
Once the dataset was prepared, I plotted the counts of successful (1) vs. unsuccessful (2) for each marketing campaign. I will show the code used:

data['campaign\_1'].value\_counts()

sns.countplot(x='campaign\_1', data=data, palette='hls')

plt.suptitle('Campaign 1 responses', fontsize=16)

plt.show()



I attach the plot for campaign 3, as this was the campaign with the highest success rate.

In doing this is I realized that there were actually 0 successful responses from campaigns 11, 12, 14 and 15; so I decided to ignore these going forward. Campaign 3 had the highest response rate so I noted that this would be probably one to watch.

I then calculated the percentage of successful campaigns using the following code:

count\_no\_sub = len(data[data['campaign\_1']==0])

count\_sub = len(data[data['campaign\_1']==1])

pct\_of\_no\_sub = count\_no\_sub/(count\_no\_sub+count\_sub)

print("percentage of no subscription for campaign 1 is", pct\_of\_no\_sub\*100)

pct\_of\_sub = count\_sub/(count\_no\_sub+count\_sub)

print("percentage of subscription for campaign 1 is", pct\_of\_sub\*100)

Campaigns 4, 6, 7, 9, 10, 13 and 16 all had a response rate of less than 1% so I decided to also not include these campaigns in my analysis, and to continue ahead with campaigns 1, 2, 3, 5 and 8.

I tried to visualise the number of subscriptions for campaign 3 grouped by job type with the following code:

pd.crosstab(data['Occupation'],data['campaign\_3']).plot(kind='bar')

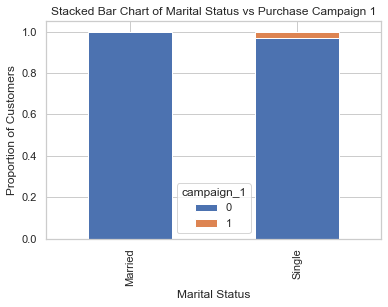
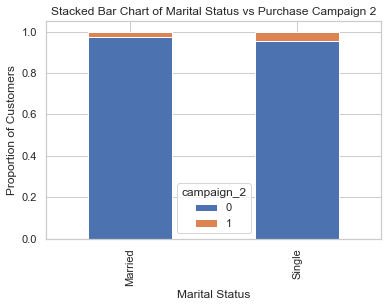
plt.title('Purchase Frequency for Job Title')

plt.xlabel('Job')

plt.ylabel('Frequency of Purchase')

plt.savefig('purchase\_fre\_job')

but the graph produced didn’t offer a helpful insight as the ratio of unsuccessful/successful is too high, and there are too many different occupations. I also plotted the same grouped by marital status which showed that marital status could be an indicator.



# 

# 

# Logistic Regression

After exploring and cleaning the data a little I wanted to create a new dataframe containing only the data I wanted to use.

# Problems with data

The campaign reaction was measured with a binary, which didn’t allow me to understand which campaigns each person has been exposed to, so I made the assumption that every person had been exposed to every campaign, which may not be true. I also couldn’t gauge which campaign has brought in the most money, this data might have been in there, but as the headings weren’t clear I couldn’t be sure.

The data also presents a challenge because there are many N/A and nan values.

The headings for the data fields were not self-explanatory so I could not understand what they all meant.

There were 16 different campaigns which is rather a lot.