Bank Marketing Campaign Report

Introduction

Problem Statement

There has been a revenue decline for the Portuguese bank and they would like to know what actions to take. After investigation, they found out that the root cause is that their clients are not depositing as frequently as before. Knowing that term deposits allow banks to hold onto a deposit for a specific amount of time, banks can invest in higher gain financial products to make a profit. In addition, banks also hold better chances to persuade term deposit clients into buying other products such as funds or insurance to further increase their revenues. As a result, the Portuguese bank would like to identify existing clients that have higher chances to subscribe for a term deposit and focus marketing effort on such clients.To resolve the problem, we suggest a classification approach to predict which clients are more likely to subscribe for term deposits.

Background

Founded 1876 in Lisbon, Obrigado Bank is the largest of the banks in Portugal in terms of total assets, and the country's largest public-sector banking corporation. It also ranks 109th on the list of major banks in the world and is the 69th largest European bank. The bank operates through branches, representative offices, and direct equity interest in local financial institutions in 23 countries located on four continents. They have contracted our services to deliver a capable model(over 80% AUC, FP<.10) of classifying whether a customer will subscribe to a new loan

Goal

Our goal is to create and implement a machine learning model that will help target customers that are more likely to subscribe to a loan. The model will have an AUC score higher than 90% and an FP score less than 10%.

Data

We received our data from the UCI Machine Learning Repository. It is a multivariate data set, with 45,211 instances and 17 attributes. The data originated from a portuguese banking institution, where they wanted to measure and improve their direct marketing campaigns.

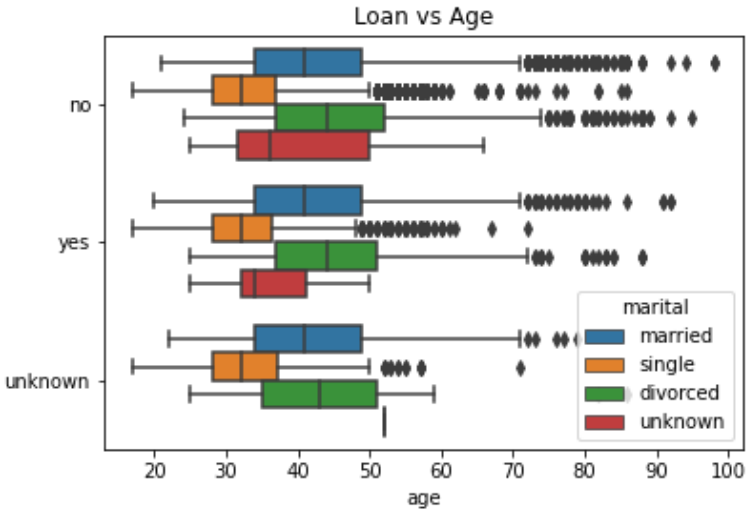
Data Cleaning and Data Wrangling

There was not much to do as far as cleaning and manipulating the data was concerned. All the data came organized and did not need to be manipulated. The biggest things we had to do was balance our data set by undersampling which we will speak more to later in the report.

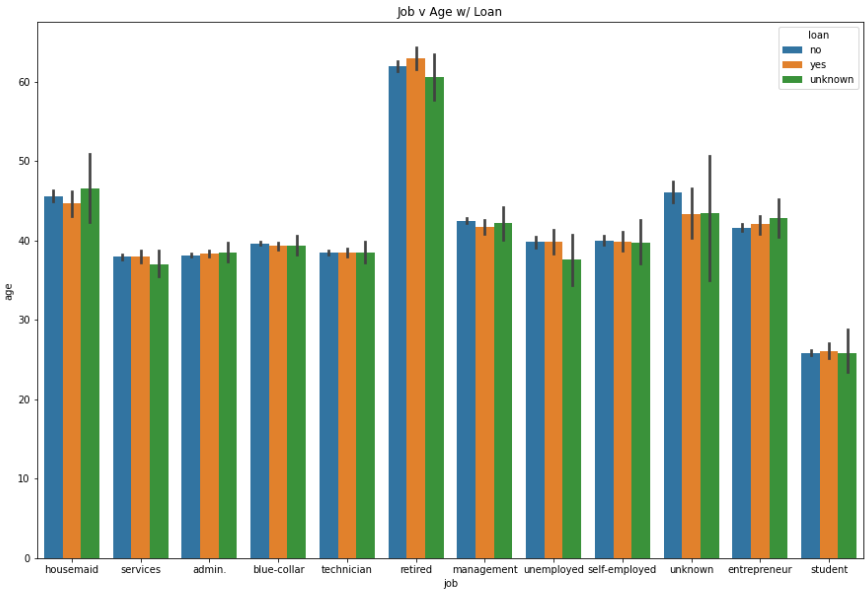
EDA & initial insights

We decided to ask some questions of the data to help guide our exploratory data analysis. For example, the questions we asked were:

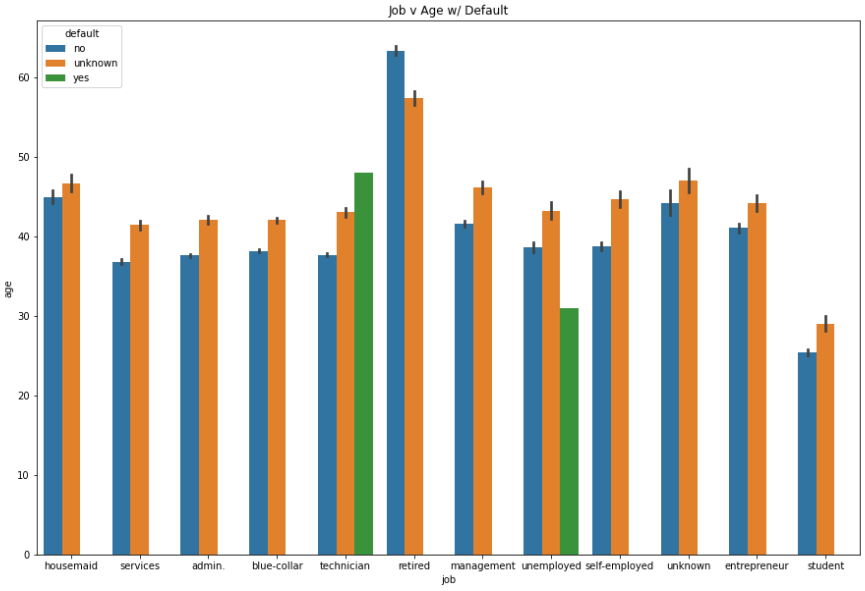
Does being married play a factor in whether a person has a loan or not ?



Does having a specific profession or career factor in whether someone has a loan?

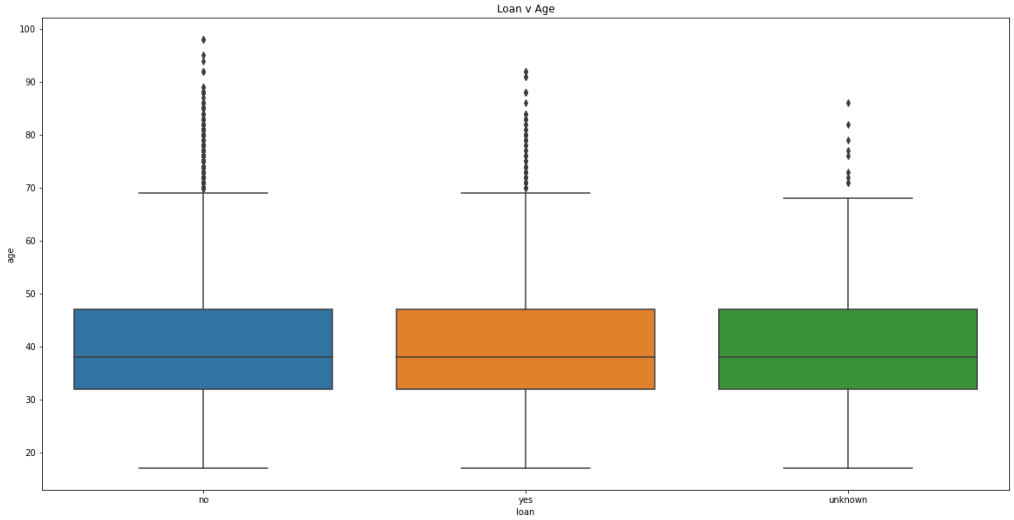


The graph above shows the comparison between occupation and likelihood to have a loan. As we can see, retired people are the highest group with current loans which makes sense. It takes 15-30 years to pay off a loan. No surprise that students are last but the standout in this graph is the amount of housemaids that have applied for and received a loan.

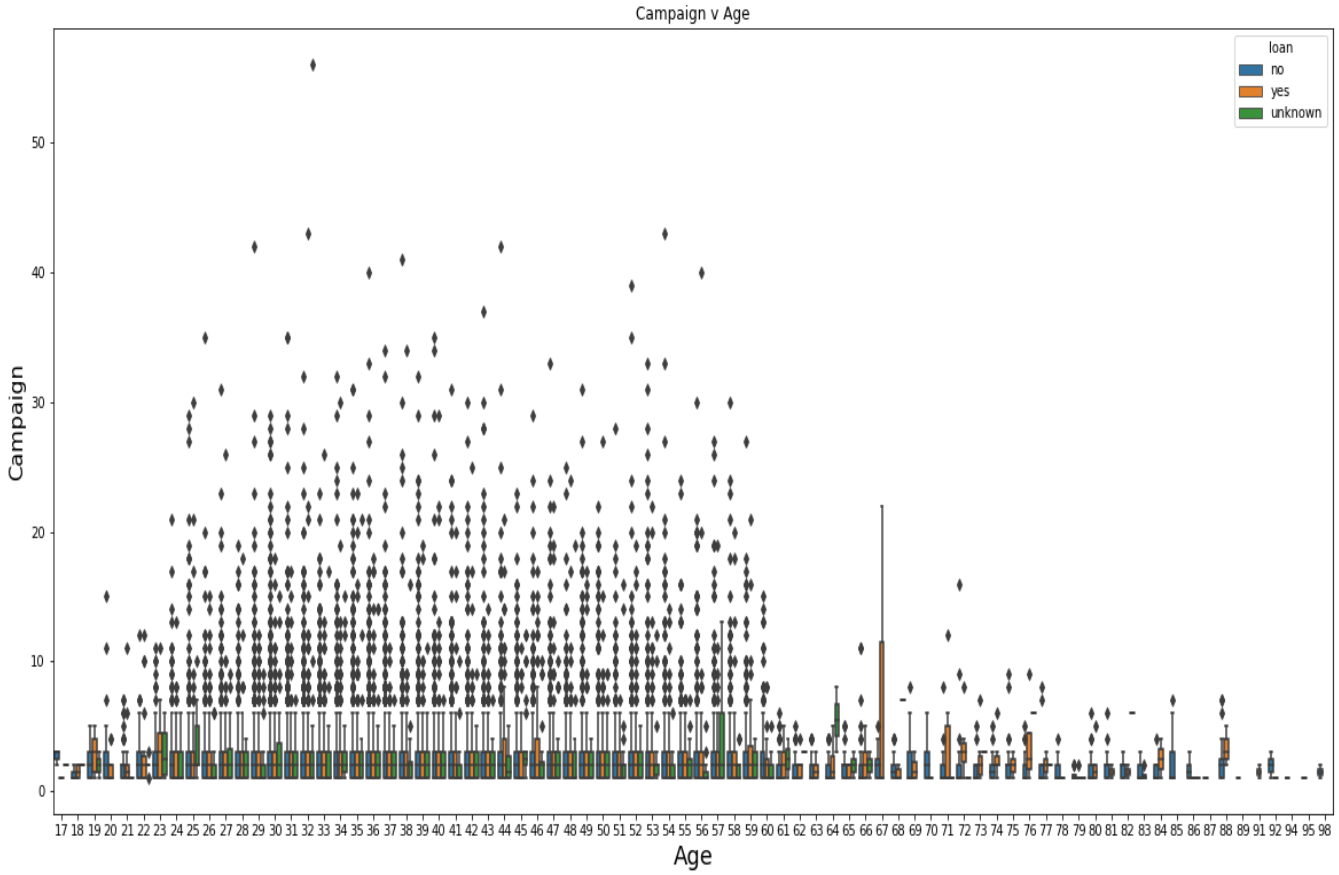


Above we can see that technicians and unemployed people have defaulted on their loan in the past. According to investopedia.com, “Default is the failure to repay a debt, including interest or principal, on a loan or security”. What is interesting to see is that technicians have defaulted and no other profession has a record of defaulting. It is concerning to not see other defaults in other categories.

Does age play a big factor in who will take a loan ?

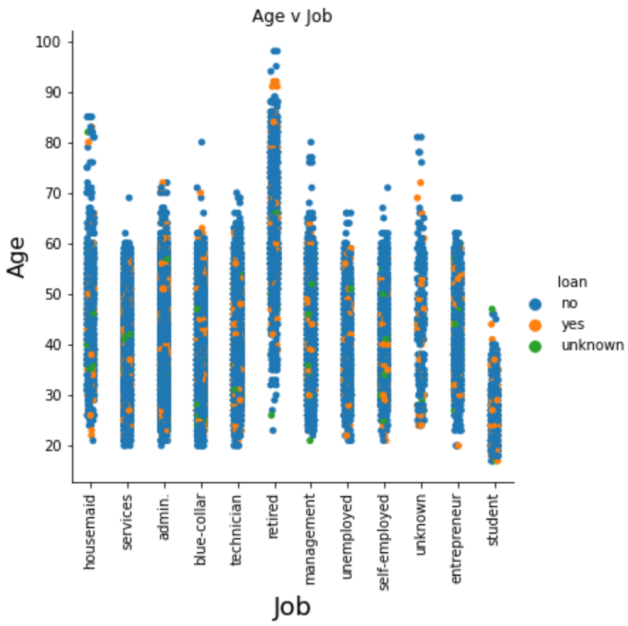


Age 40-50 is the range of people that responded yes to credit default. This age range has the most risk of defaulting on their loan.



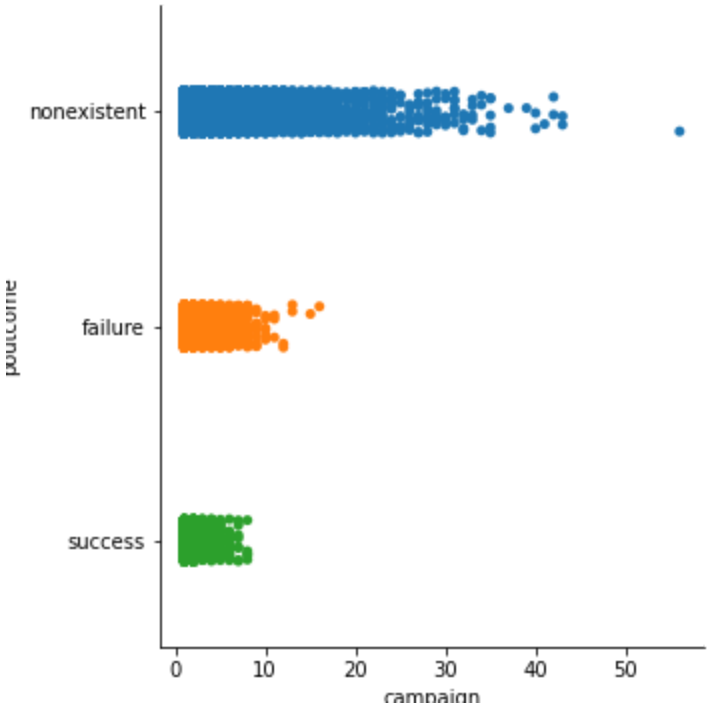
This graph shows the ages that said yes or no to a new loan. Not including the outliers, the campaign was pretty successful within 10 days if they were able to get in contact with their prospects.

Age 19,23,44,46,59,67,71,76 are the ages where we see increased spikes in subscriptions. Based on the age ranges, you could create custom campaigns to target each age segment.

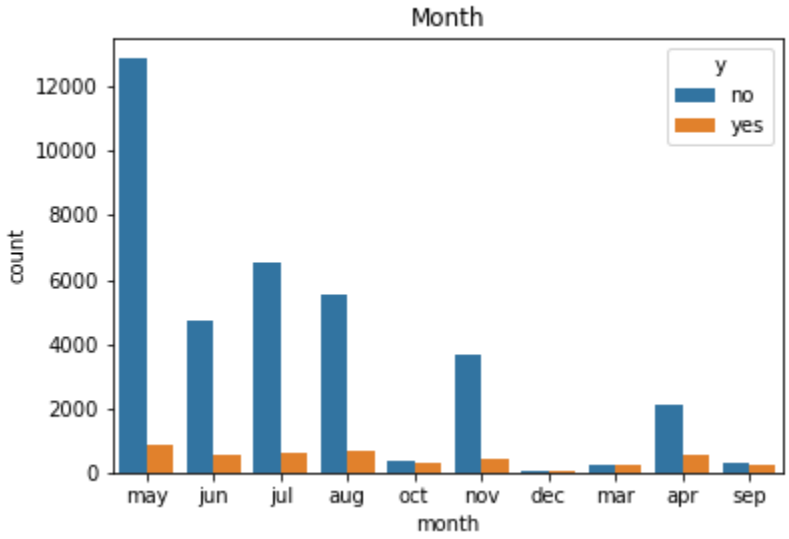


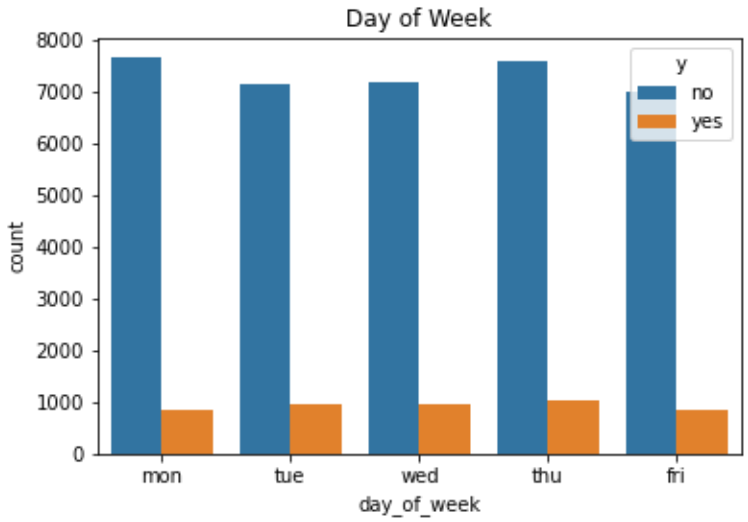
Age is definitely a factor, we can see a lot of people that have loans in their 90's which is very off putting. The loans are very specific to where their respective occupations interact with age. For example,we can see lots of retired people with loans at a higher age than you would see a housemaid getting a loan or a blue collar worker. This suggests that lifestyle marketing could be an avenue to more specifically target their clients. "Lifestyle segmentation is the process of dividing a market of potential buyers into different groups of people with similar ways of living. "(The balancesmb.com).

How did the previous campaign perform ?



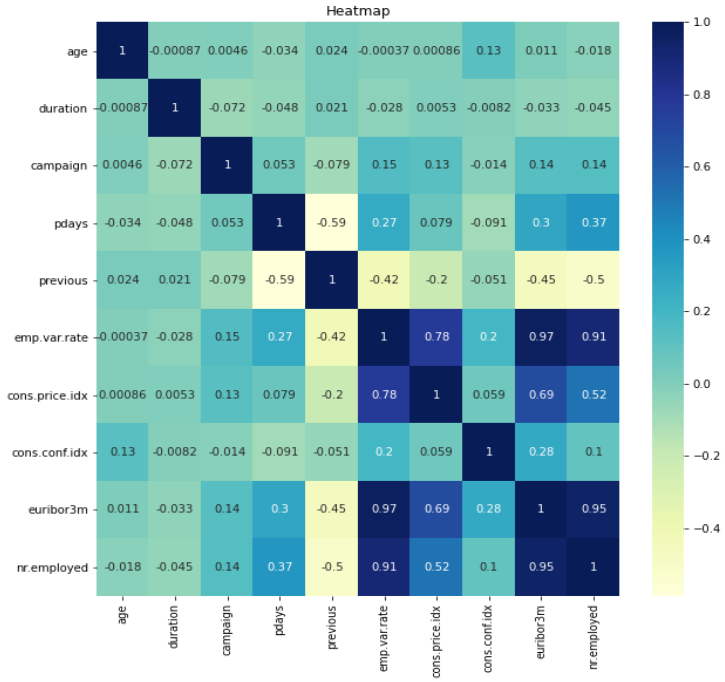
The highest chance of success for the campaign depended on employees being able to perform and execute calls within 10 days of starting their marketing campaign. There is not an outlier to support making attempts after 10 days.





The previous campaign resulted in 4640 people saying yes and 36,548 people saying no. That comes out to just over 11% of people saying yes to a loan. Depending on what constitutes a successful campaign, from an outside perspective it did not behave poorly however, it can be revamped to work more efficiently. In addition, as we can see from the day of week graph, the best times to call during the week are Tues,Wed, and Thursday. Also, the best time of the year for the campaign seems to be the Spring and Summer months, with the peak being in May.

Heatmap



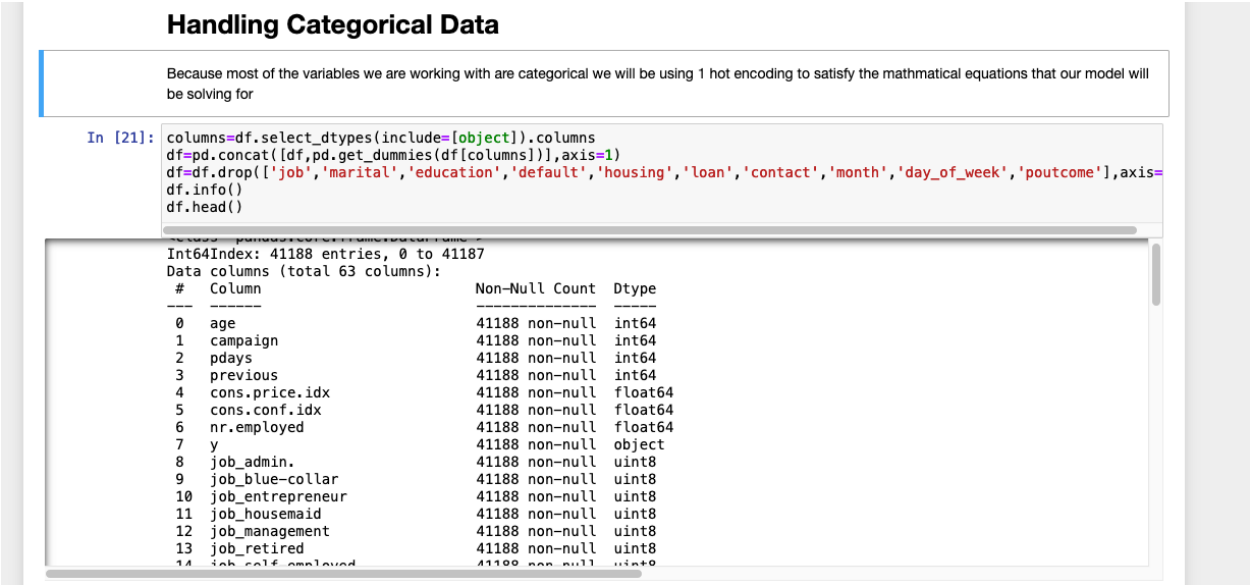
From the heatmap, we can see that employees play their part whether it's the number employed or their variation rate. In addition, it seems like pdays and previous have a correlation as well to how likely someone will subscribe to a loan.

In summary, from our Eda we were able to find: relationships to explore such as pdays and previous with our target variable. April through August were the best months of the campaign, outperforming other months significantly. Best time of the week to call customers are: Tuesday, Wednesday, Thursday. There is not enough information to say whether to target one profession over another but there is added risk in evaluating unemployed and technicians who apply for a loan vs the rest of the other occupations. Also, married people are most likely to say yes to a loan but that's just because there were so many married people polled vs single people. Lastly, age does affect our target variable. The younger someone is, specifically around 18 but the range can be as far as 60 years old, the more likely they will request a loan.

Preprocessing

The first preprocessing step that we took was to remove any feature with a correlation of > 80%. Because in the correlation matrix, we saw how highly the company's internal statistics were such as employee variation rate.

<u>Transforming categorical data:</u>
The next step we applied was using one hot encoding on our features. The reason why is because we have no particular order that our features need to be encoded in, which makes choosing one hot encoding the obvious choice.



Handling Categorical Data

Because most of the variables we are working with are categorical we will be using 1 hot encoding to satisfy the mathmatical equations that our model will be solving for

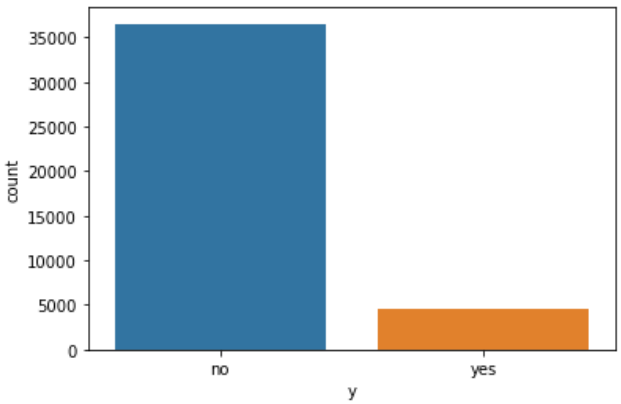
In [21]:
```
columns=df.select_dtypes(include=[object]).columns
df=pd.concat([df,pd.get_dummies(df[columns])],axis=1)
df=df.drop(['job','marital','education','default','housing','loan','contact','month','day_of_week','poutcome'],axis=
df.info()
df.head()
```

```
Int64Index: 41188 entries, 0 to 41187
Data columns (total 63 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   age                            41188 non-null  int64
 1   campaign                       41188 non-null  int64
 2   pdays                          41188 non-null  int64
 3   previous                       41188 non-null  int64
 4   cons.price.idx                 41188 non-null  float64
 5   cons.conf.idx                  41188 non-null  float64
 6   nr.employed                    41188 non-null  float64
 7   y                              41188 non-null  object
 8   job_admin.                     41188 non-null  uint8
 9   job_blue-collar                41188 non-null  uint8
 10  job_entrepreneur               41188 non-null  uint8
 11  job_housemaid                  41188 non-null  uint8
 12  job_management                 41188 non-null  uint8
 13  job_retired                    41188 non-null  uint8
```

Taking care of unbalanced data

The last preprocessing step to take care of is the imbalance that will invariably create noise if we do not remove it right now. The process that we originally used was oversampling, however, most models that were used did perform or even meet our expectations. Considering that we only have 8GB of ram to use and diminishing gains, we decided to try another approach. Instead, we did the opposite by oversampling, using the smote technique which creates synthetic data. At this point of the project, we do have to declare that we have introduced a new form of bias in the form of synthetic data. This may affect our ability to generalize with our model moving forward.

Before



After Oversampling

```
Before oversampling:  Counter({'no': 36548, 'yes': 4640})
After oversampling:  Counter({'no': 36548, 'yes': 36548})
```

This is the result we were after.

Modeling

Before modeling we decided to split the data using test train split. The reason why we want to do that is so we can have the ability to generalize our results without having to worry about bias.

Something to also point out is in every model we used we have incorporated some form of cross validation as well as hyper parameter tuning. We have done so to increase the performances of our models as well as increase our ability to reduce as much bias/overfitting as possible.

Time for the best part! We will begin our analysis of the models we chose, first beginning with KNN. according to towarddatascience.com, “KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).”

Our result were(KNN):
* FP 37%,FN 21%, AUC 75% with undersampling
* FP 7.9% FN 7.3%, AUC 96% with oversampling

After Knn, we decided to try logistic regression out. Mostly because it's a simple model that works efficiently and may give us a great starting point. We like using LogisticRegressionCV because there is L1&L2 regularization applied under the hood that we don’t have to add.

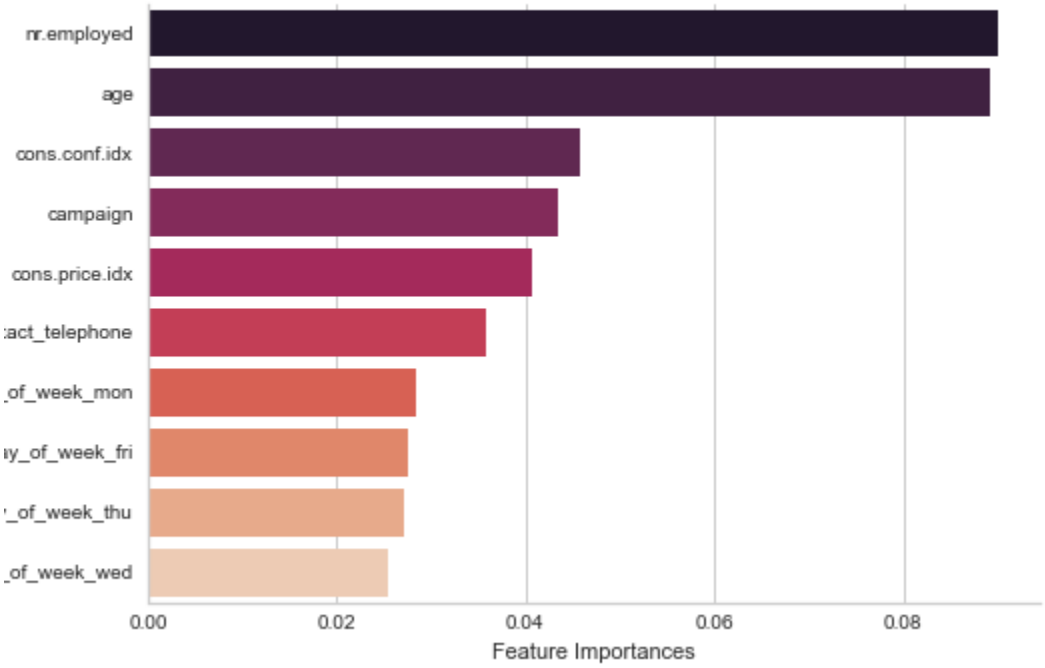
That being said, our results were(Logistic Regression):
* FP 38% FN 14% AUC 80% with undersampling
* FP 11% FN 1.9%, AUC 97% with oversampling

Next on the list is our instance of using SVM. Although it takes a lot computationally when it is firing correctly, it works very well to create a line or a hyperplane which separates the data into classes.
Our results were (SVM):
* FP=29%,FN=32%,AUC=74% with undersampling
* FP=11% FN=2.3% AUC=97% with oversampling

After SVM, we decided to use Random Forest. The reason why is because it is so versatile and accurate when combined with cross validation. In addition, the feature importance list that Random Forest generates is worth its weight in gold for the information that it provides. Especially considering the nature of our classification problem.
The results were(RF):
* FP=37%,FN=13%,AUC =80% with undersampling
* FP= 7.1% ,FN=5.2%, AUC=98% with oversampling

The last model we wanted to try was XGBoost. The reason why is because XGBoost is because of its ability to handle missing values and prevent overfitting. One of the main things that contributes to XGBoost's prowess is the regularization that this model offers. Our results were better than we were expecting (XGBoost):
* FP=36%,FN=18%,AUC=78% with undersampling
* FP=9.2% ,FN= 2.9%, AUC=97% with oversampling

Overall, the model that we have chosen to implement is Random Forest. The reason why we chose it is because of the AUC accuracy, in addition to having one of the lowest type 1 error rates,which is more important for us than type 2 error in our study. The reason being, the campaign would be negatively affected more by false positives than false negatives because if we have too much type 1 error, we really don't know how well the campaign is performing due to bias.



It looks like employees play the biggest factor with how many employees are there to make the calls.Volume is another factor, because we can see how important the number of employees to carry through the campaign is.Any decrease in the workforce suggests that would have negative effects to the campaign. Age also seems to be a factor,but that is no surprise because the younger they can get a client, the longer they can do business with them over the course of their life.Consumer confidence plays a factor due to its relationship with consumers and how well they believe the economy is doing. The

better the economy the more likely the customer will take a loan because a good economy equals stable work for most.

Takeaways

What we learned

From our EDA, we learned that calling during the middle of the week gave our client the highest chance for success. In addition, April-August produces the highest results for the campaign. That being said, Age is a factor as we can see from the feature importances chart and heatmap. After evaluating the poutcome with regard to the previous campaign, it would be safe to say that any chance of success will happen within 10 days of direct marketing to a customer; after 10 days there were no successful attempts. Another point to add is the age groups that subscribed to a loan can be segmented in such a way that can help target consumers of the same age for the next campaign. It looks like employees carry high significance, with regards to the feature importances chart, with how many employees are there to make the calls.Volume is another factor, because we can see how important the number of employees to carry through the campaign is.Any decrease in the workforce suggests that would have negative effects to the campaign.With regards to modeling,Logistic Regression gave us the lowest FN rate. Random Forest gave us one of the highest AUC scores as well as FP score less than 10%.

Recommendations to improve campaign

The recommendations that we believe will positively affect the next campaign are:

- Focus on calling customers Tues,Wed,Thur for highest success
- Drop any customer that hasn't subscribed in at least 15 days or 8 call attempts.
- Keep employees working and happy, any drop in workforce affects the campaign.
- Segment customers by age and incorporate lifestyle segmentation as well with regards to occupations and how they vary by age. Ex: A 40 year old executive should be marketed differently than a 22 year old college graduate.
- Focus on targeting entrepreneurs, through specified business loan offerings and deals. In addition, target the management occupation in order to help drive volume.

Future Work

We would have liked to incorporate time into our analysis. That way we could see how the campaign performed over time. In addition, adding more data would only make our

model more robust. What we specifically mean is adding more data like balance in the customers bank account, their wages and debt(debt to income ratio). Also perhaps some geographical information to see if we could cluster customers based on location.

https://github.com/astephens10/Final_project

References

1. https://www.academia.edu/6412064/A_Data_Driven_Approach_to_Predict_the_Success_of_Bank_Telemarketing