

Bank marketing Analysis using Machine Learning



#DHEERAJ PRANAV
B. Tech,CSE

Understanding the problem

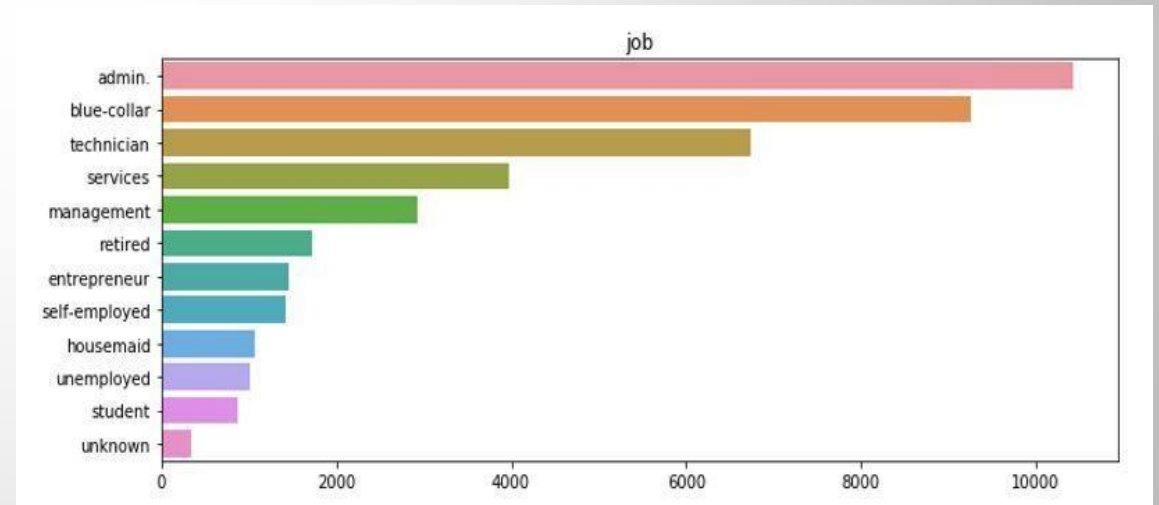
- Problem Statement: Improve marketing campaign of a bank by analyzing their past marketing campaign data and recommending which customer to target
- Problem Motivation: By devising such a prediction algorithm, the bank can better target its customers and better channelize its marketing efforts
- Banks offered their clients fixed-term products such as CDs. Data was collected about each client, type of contact, and outcome.
- What can this data tell us about marketing success for this campaign?
- Can these data science techniques be applied to other areas?

#SOURCE OF DATA

<https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>

Data Exploration

- All coding is done in Python 3.
- Extensive use of pandas, numpy, matplotlib, as well as seaborn and sklearn packages.
- Dataset contained 19 different features on more than 41,000 clients.
- Features were both categorical and numerical. Target variable was binary ("Yes" or "No").
- Pandas package was imported and a dataframe was created.
- Categorical variables were looked at first. Visualizations were created using the seaborn package.



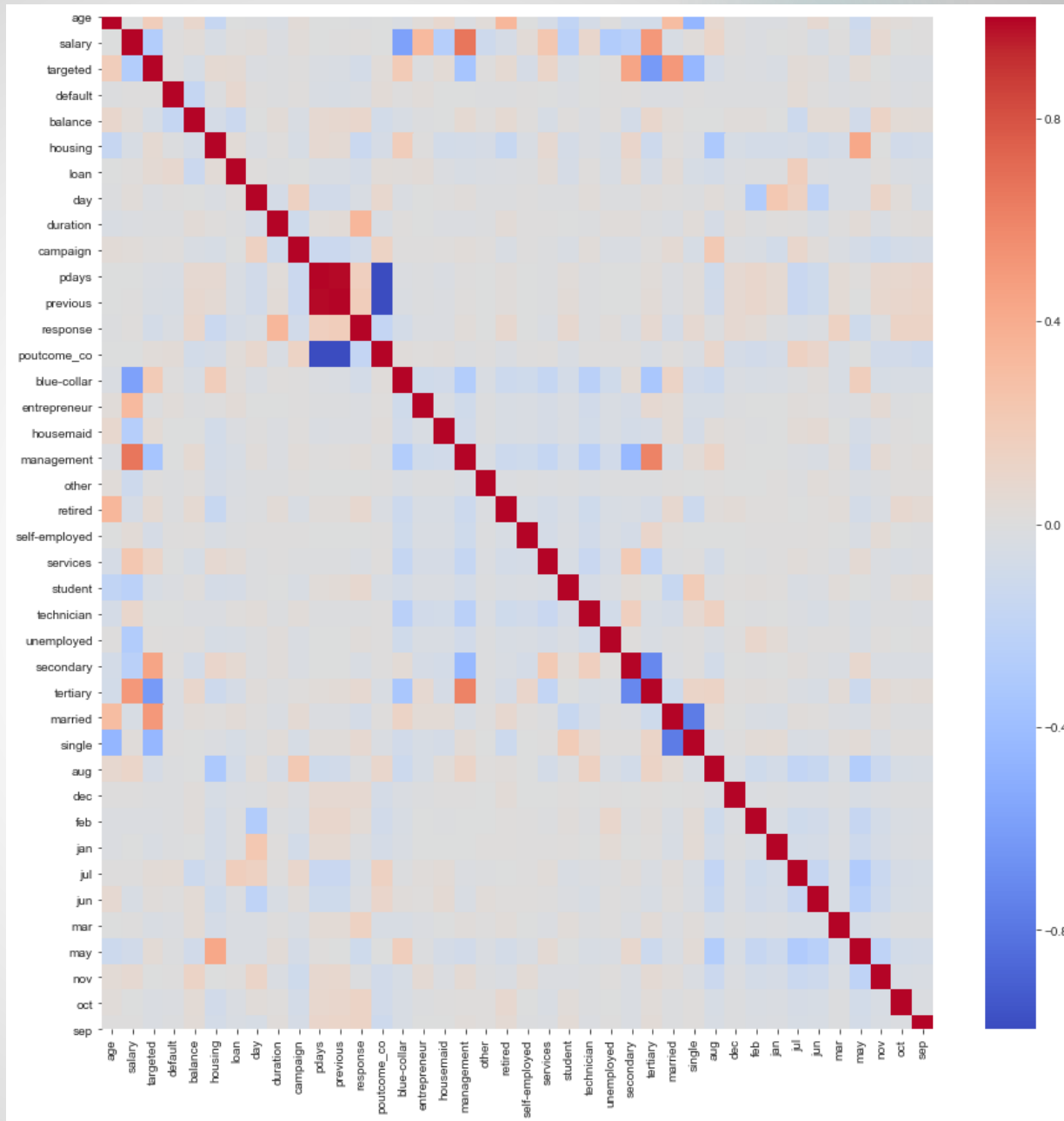
#DRIVE LINK OF EXPLANATION

<https://drive.google.com/drive/my-drive?tid=0By9jKmOK17EVZmxsZ1BBX2J3enM>

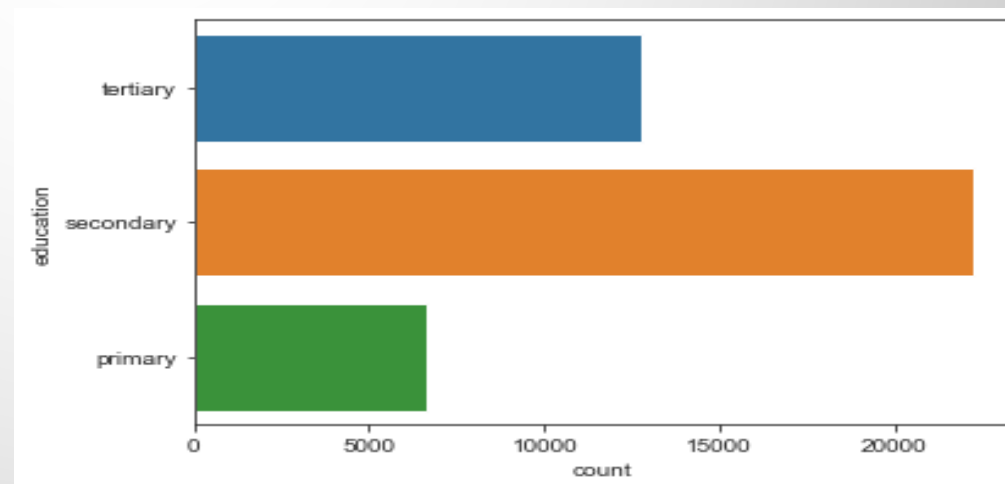
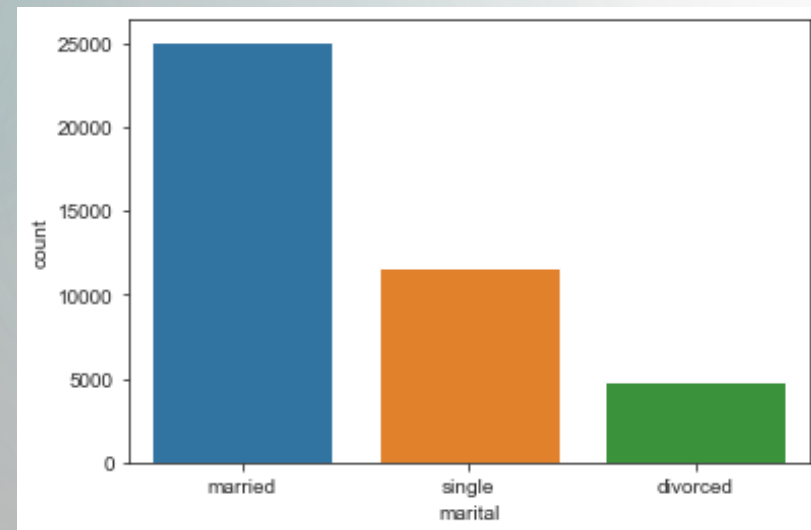
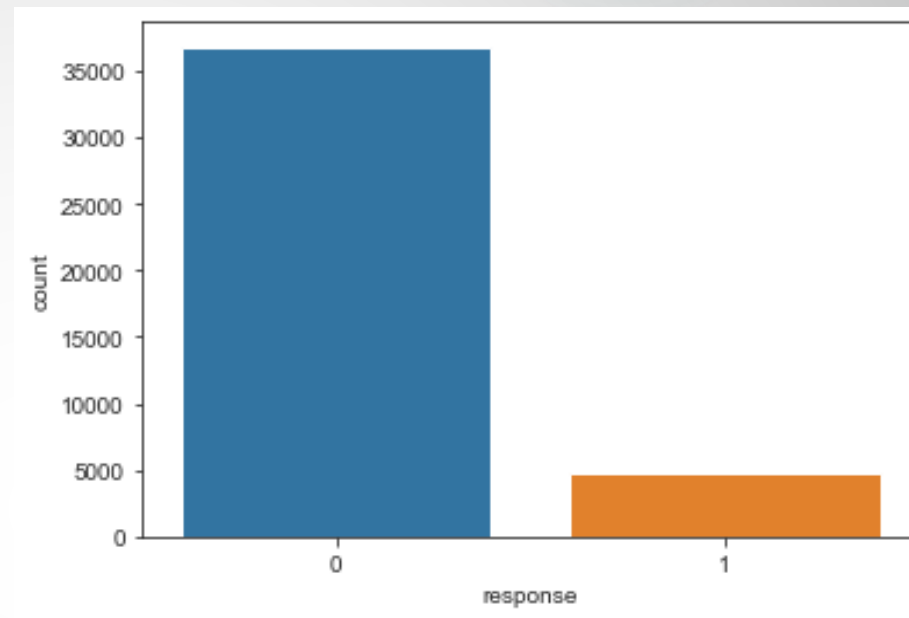
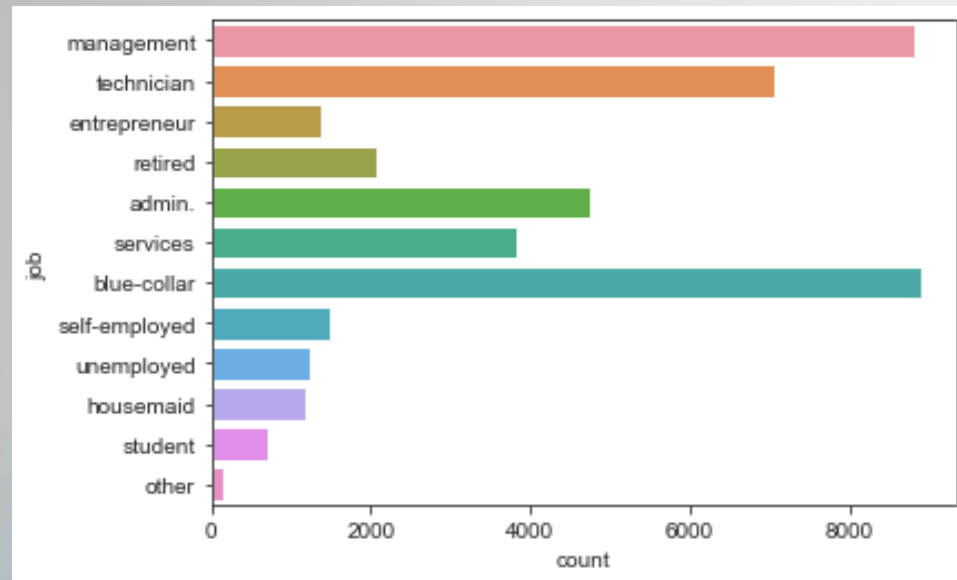
Data Preperation

- Many features had missing values. How do we handle this?
- For categorical features, imputation using other independent variables. For example, cross-tabulation between 'job' and 'education'; 'age' and 'job'; 'home ownership' and 'loan status.'
- Among numerical features, fortunately only column ('pdays') had any missing values. Unfortunately, missing values made up the majority of the column.
- To handle this, 'pdays' was converted from a numerical feature to a categorical feature using buckets: < 5 days, 6-15 days, etc.
- Heatmap using seaborn package was created to show us any particularly strong correlations between the independent variables and the target variable outcome.

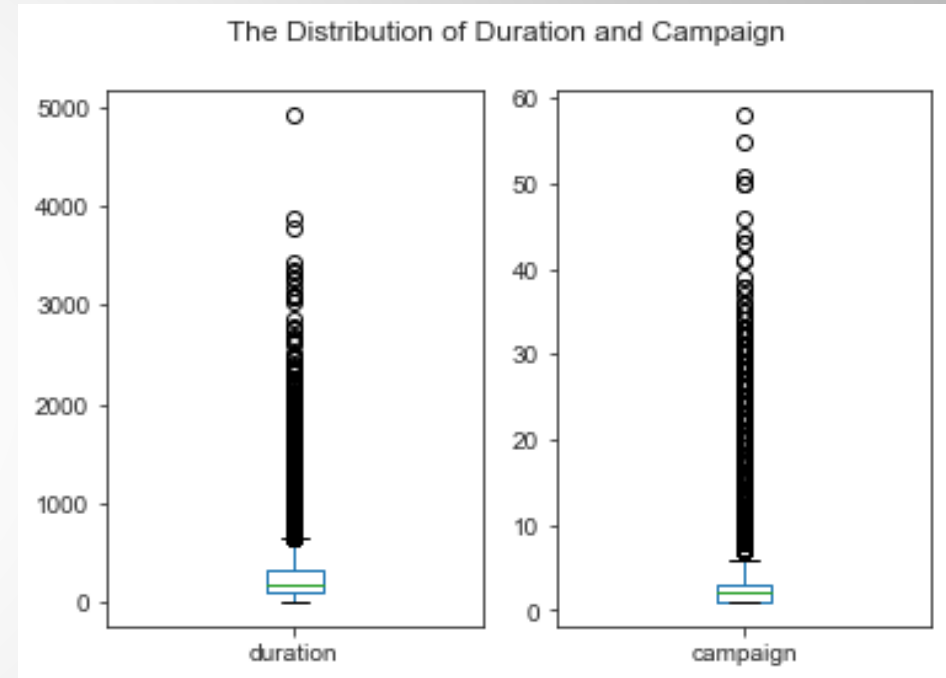
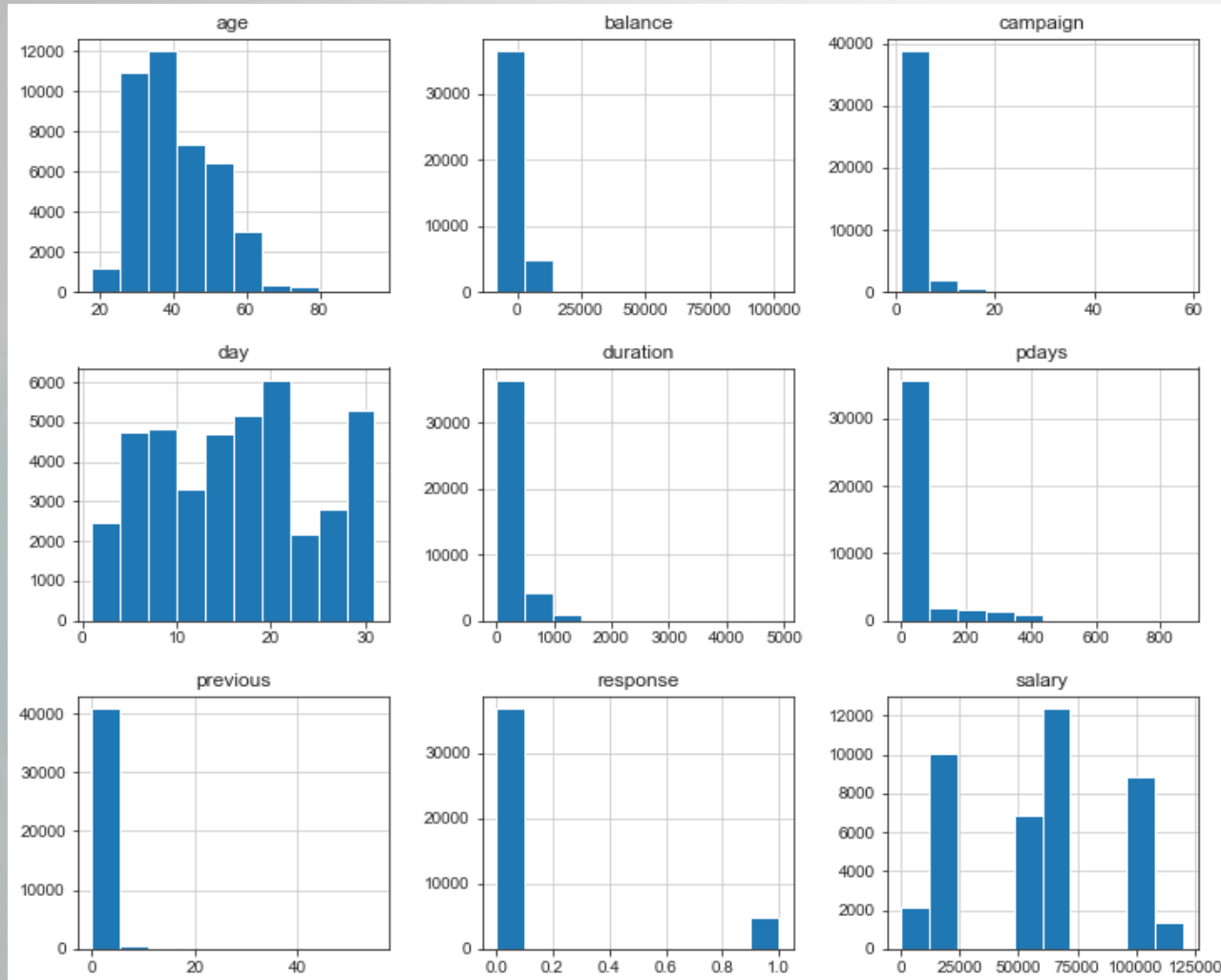
Correlation Heat map:



Plots for Categorical Data



Visualizations for Numerical Data



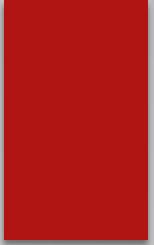
Model Building

Logistic Regression

- `sklearn.linear_model.LogisticRegression`
- Its a classification model though name is Logistic regression
- Fits a sigmoid function to a data
- Outputs probability which is in $[0,1]$ range unlike linear models.

Random Forest

- `Sklearn.ensemble.RandomForestClassifier`
- Constructs multiple decision trees and takes the mode of those trees for an example to make the final decision
- Individual Trees are intentionally over fit and validation set is used to optimize the forest level parameters

- 
- ❑ The dataset is divided into training data and test data.
 - ❑ With the intention of using the training data to find the parameters of the particular model being used (fitting the model on the training data) and then applying this to the test data to determine the model's performance .
 - ❑ And to draw conclusions about its predictive capability.
 - ❑ This can be done with a `sklearn.cross validation.train test split` function call by specifying split ratio.

Feature Selection

➤ Significant features are detected and selected using various methods during model building:

1. Recursive Feature Elimination
2. Variance inflation factor
3. OLS (p values)

➤ Significant features are:

['poutcome_co',
'student',
'month_dec',
'month_mar',
'month_may',
'month_oct',
'month_sep']

Out[983]:

	Features	VIF
0	poutcome_co	1.39
4	may	1.36
1	student	1.02
3	mar	1.01
5	oct	1.01
6	sep	1.01
2	dec	1.00

Model Evaluation

- Model's of different types of algorithms are evaluated based on different metrics , here we have used :
 1. Accuracy Score
 2. Precision and Recall
 3. K-Fold (cross validation score)
- logistic regression achieved an accuracy of about 88%, suggesting a high level of strength of this model to classify the customer response given all the defined customer features.
- Random forest has also got somehow similar accuracy.

Accuracy Measures

1. Accuracy score for selected features when we use logistic model is

0.8923

and when we use Random forest model is

0.8915

2. Precision score for selected features when we use logistic model is

0.5510

and when we use Random forest model is

0.5204

3. Cross val score for selected features when we use logistic model is

0.88557

and when we use Random forest model is

0.8855

#Confusion Matrix

Out[787]:

	predicted no	predicted yes
actual No	7207	200
actual yes	606	303

CONCLUSION

According to previous analysis, a target customer profile can be established. The most responsive customers possess these features:

- Feature 1: age < 30 or age > 60
- Feature 2: students or retired people
- Feature 3: specific months (dec , mar , may ,oct)

By applying logistic and Random forest classification algorithms estimation model were successfully built. With these two models, the bank will be able to predict a customer's response to its telemarketing campaign before calling this customer. In this way, the bank can allocate more marketing efforts to the clients who are classified as highly likely to accept term deposits, and call less to those who are unlikely to make term deposits.

#Recommendations

- Try to engage customers and have longer calls
- Preferably use Telephone as the mode of contact
- Prioritize those customers to who were part of the previous marketing campaigns.

Thank You.