Data Scientist for Enterprise

Nano Degree Program

Capstone Project Report

By: Durga Parida

12/9/19

Definition

Project Overview

Data Analytics is a fast-growing field in competitive business world. Businesses are aiming to transform Big Data into actionable intelligence by leveraging AI (Artificial Intelligence) and ML (Machine Learning).

The data driven decision making is becoming an integral part of Company's core strategy to derive the customer insights for making better business decisions. Businesses are using the ML models to develop the prediction models in areas such as Operational improvement, i.e. Customer Retention, Fraud Prevention, Price Modeling etc.

The goal of this capstone project is to design and deploy a Supervised ML "Binary Classification" model, such as Logistic regression to predict if a Bank customer will subscribe(yes/no) a term deposit campaign (Response variable). This will help to execute some prescriptive measures (i.e. revised campaign strategy to offer better interest rate/promo etc.) to increase the term deposit acceptance rate.

Problem Statement

I have used the data is which is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe (yes/no) a term deposit (Response variable).

This dataset is sourced from the **IBM Watson Platform**- UCI Machine Learning Repository. This data set contains 10% of the examples and 17 inputs, randomly selected from the full data set, bank-full.csv. The full data set is available on the Watson Studio Community as well as at https://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (Response variable).

Metrics

Grid Search Cross validation- I have used the Grid Search Cross validation to measure the effectiveness of the model. This step randomly divides the set of observations into k groups. The 1st set is treated as Validation set and the method is fit (Training) on remaining k-1 fold

In this project, I have used CV = 10, i.e. 10- fold cross-validation. Basically, we could have evaluated the score in a loop for different values of CV to get the highest score. I have used CV = 5 and CV =10. Both yielded around **87%** accuracy. So, I have decided to use 10-fold validation

```
Fitting 10 folds for each of 40 candidates, totalling 400 fits
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 8 concurrent workers.
[Parallel(n_jobs=-1)]: Done  34 tasks      | elapsed:    9.6s
[Parallel(n_jobs=-1)]: Done 184 tasks      | elapsed:  2.4min
[Parallel(n_jobs=-1)]: Done 400 out of 400 | elapsed:  3.2min finished
best_score(Mean cross_validated score) : 0.8712766271499693
Best Penalty: l2
Best C: 2.7825594022071245
Best Parameters: {'C': 2.7825594022071245, 'class_weight': 'balanced', 'dual': False,
'fit_intercept': True, 'intercept_scaling': 1, 'l1_ratio': None, 'max_iter': 100, 'mul
ti_class': 'warn', 'n_jobs': None, 'penalty': 'l2', 'random_state': None, 'solver': 's
aga', 'tol': 0.0001, 'verbose': 0, 'warm_start': False}
```

Choosing the right Metric for Model evaluation

The selection of the correct metric to evaluate the model depends on the business objective , the problem we are trying to solve and the Model we are using.

There are many supervised machine learning models are available today, such as:

- **Prediction**
  - Linear Regression
  - K-nearest Neighbor
  - Regression Trees
  - Neural Networks etc.
- **Classification**
  - K-Nearest Neighbor
  - Naïve Bayes
  - Classification Trees
  - Logistic Regression
  - Support vector Machines
  - Neural Networks etc.
- **Time Series Forecasting**

Since I have selected Logistic regression for our Project, let's analyze various metrics that can be used to evaluate and validate our model. We know that we can't use the Metrics that are used for Regression Models (such as AIC, BIC, , Mean Squared Error, R^2 etc.)

Common metrics that are used for evaluating Classification Models

- *Classification Accuracy* – It applies to the data set when there are an equal number of observations in each class and that all predictions and prediction errors are equally important, which is not the case either.
- *Log Loss*- Logistic loss (or log loss) is a performance metric for evaluating the predictions of probabilities of membership to a given class.
- Area Under ROC Curve- performance metric for binary classification problems.
- *Confusion Matrix*- The confusion matrix is used to describe the performance of a classification model on a set of test data for which true values are known.
- *F1 Score*- It is the weighted average of the Precision and the recall scores. The F1 score reaches its perfect value at one and worst at 0.It is a very good way to show that a classifies has a good recall and precision values.

Justification for using the following Classification Metrics for our Project

1. **Precision and Recall** metric for small positive class

   Justification: This metrics is primarily used for Classification models with especially an imbalanced data set(*i.e. # of not-accepted term deposits datapoints are higher than # of already-accepted datapoints)*

   - Precision (*Positive Predictive Value*): # True positives / # predicted positive = TP/(TP+FP).
   - **Recall (Sensitivity)**: TPR (*True Positive Rate*) = This will be our key metric as we will be measuring our performance against, since our primary goal is to correctly predict the positive cases (i.e. Customer will accept term deposit offer)
     - Recall = TP / (TP+FN)= **81%**

2. **ROC**(Receiver Operating Characteristic)

   - *When we want to give equal weight to both classes Prediction ability, we can look at the ROC curve. In our case (as stated in <Result Section>, AUC (Area Under Curve) is **88%**

   Justification: Use precision and recall addressing small positive class (*i.e. Customer will accept term deposit offer)* — When the positive class is smaller and the ability to detect correctly positive samples is our main objective, we should use precision and recall.

II. Analysis

Data Exploration

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe (yes/no) a term deposit (Response variable).
This dataset is sourced from the **IBM Watson Platform**- UCI Machine Learning Repository.

The full data set is available on the Watson Studio Community as well as at https://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

Started with 45,211 Data points and 17 features/columns

```
Data shape (rows, cols): (45211, 17)
```

Key Predictor/feature variables, but not limited to:

Continuous/Numeric Variables:

- Age
- Balance
- Campaign(# of contacts performed during this campaign)
- Duration (Last Contact duration)

Categorical Variables:

- Job
- Marital Status
- Education
- Default to a loan Payment
- Own House
- Loan taken

Target (Response variable):

- Term Deposit (Yes/No)

Sample Data (Source: `bank_marketing_data.csv (attached in zip file)`

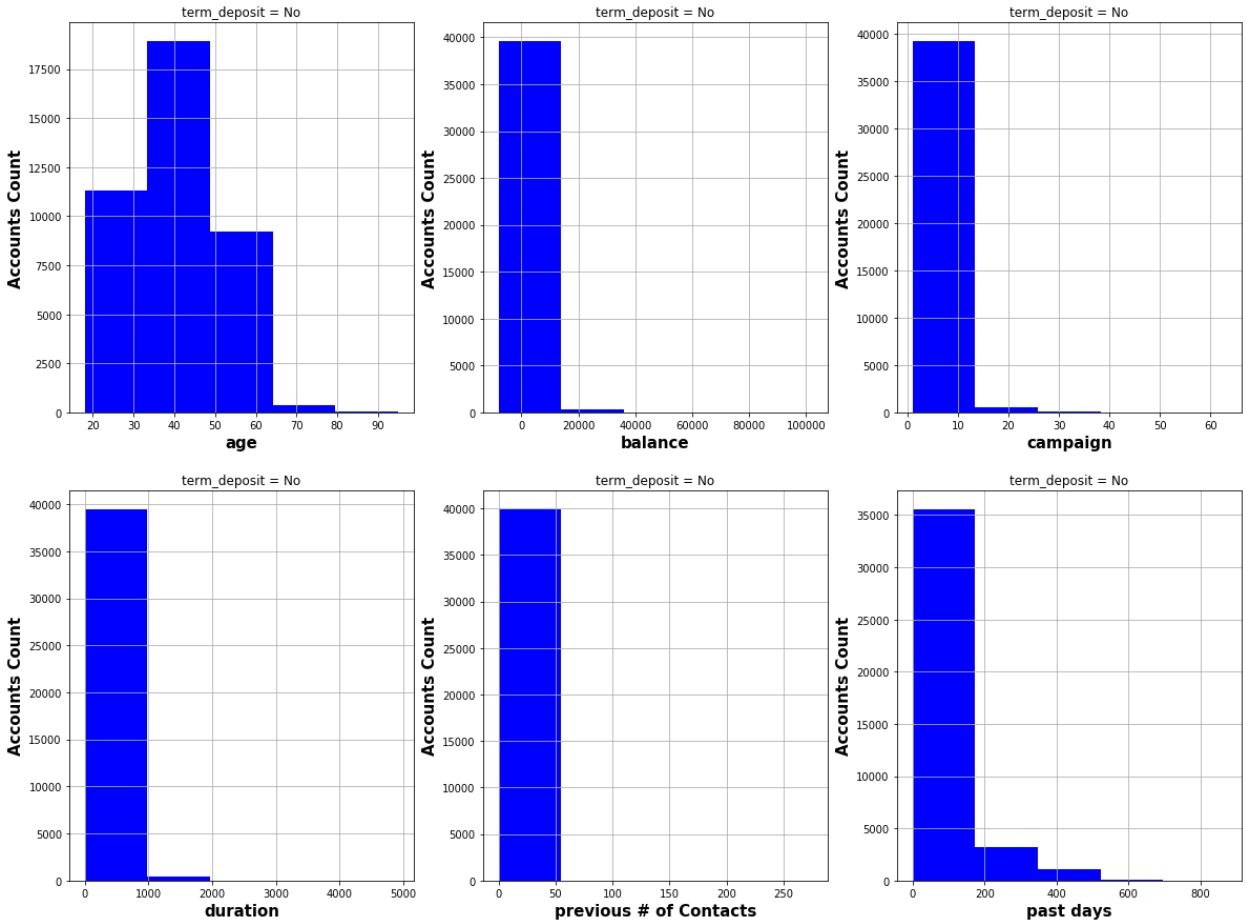
age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	term_deposit
58	anagemer	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
33	ntreprene	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
35	anagemer	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
28	anagemer	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
42	ntreprene	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
58	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no
43	technician	single	secondary	no	593	yes	no	unknown	5	may	55	1	-1	0	unknown	no
41	admin.	divorced	secondary	no	270	yes	no	unknown	5	may	222	1	-1	0	unknown	no
29	admin.	single	secondary	no	390	yes	no	unknown	5	may	137	1	-1	0	unknown	no
53	technician	married	secondary	no	6	yes	no	unknown	5	may	517	1	-1	0	unknown	no
58	technician	married	unknown	no	71	yes	no	unknown	5	may	71	1	-1	0	unknown	no
57	services	married	secondary	no	162	yes	no	unknown	5	may	174	1	-1	0	unknown	no
51	retired	married	primary	no	229	yes	no	unknown	5	may	353	1	-1	0	unknown	no

Exploratory Visualization

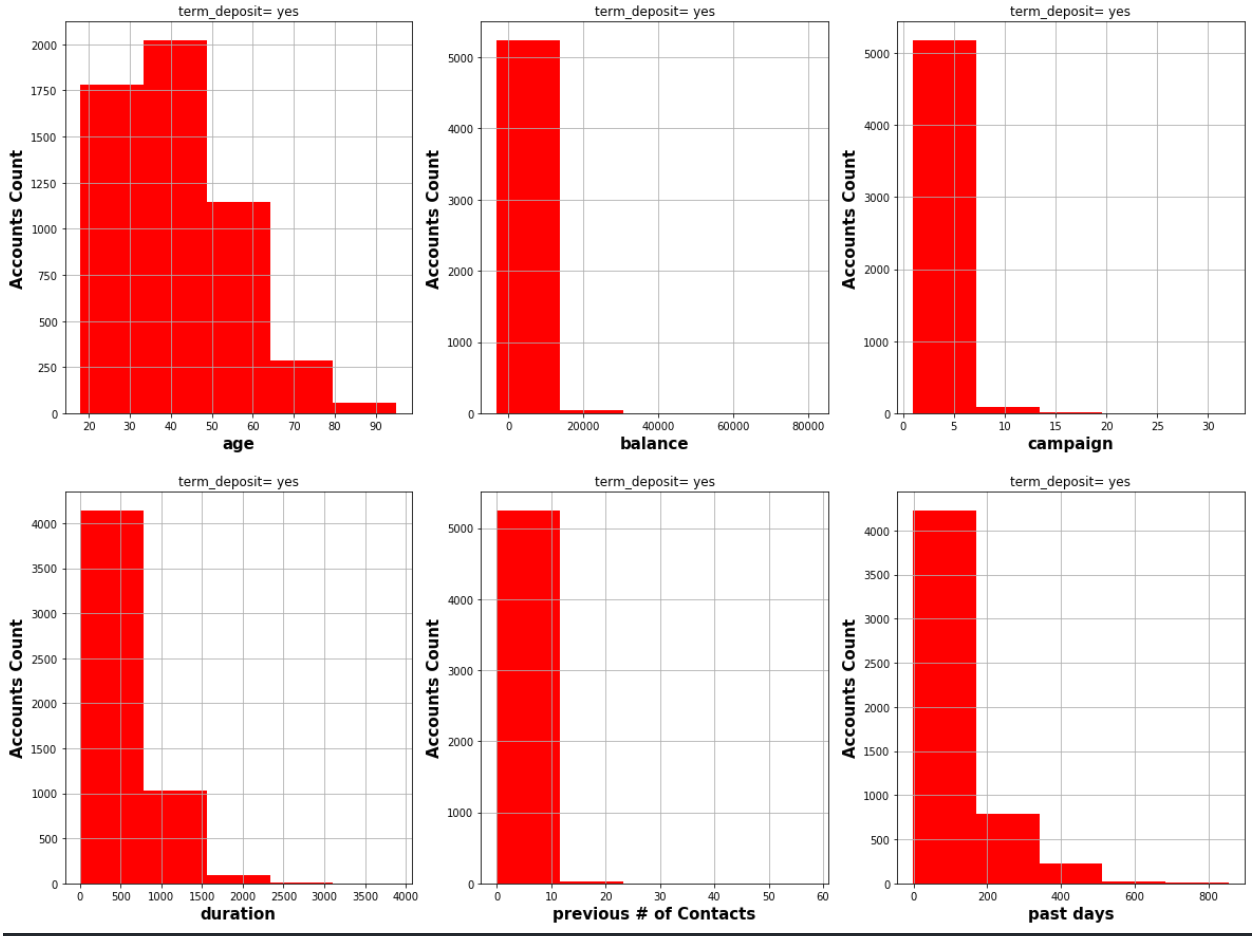
1. **For Continuous/Numeric variables** (age, balance, duration etc.):

   We have calculated central tendency and spread of variables.
   As we can see, there are some outliers present in <balance> variable with high negative credit balance. We will be removing the outliers as then part of Data Pre-processing step using *z-score* method

   Histogram for Numeric Variable with respect to term deposit Flag = "No"



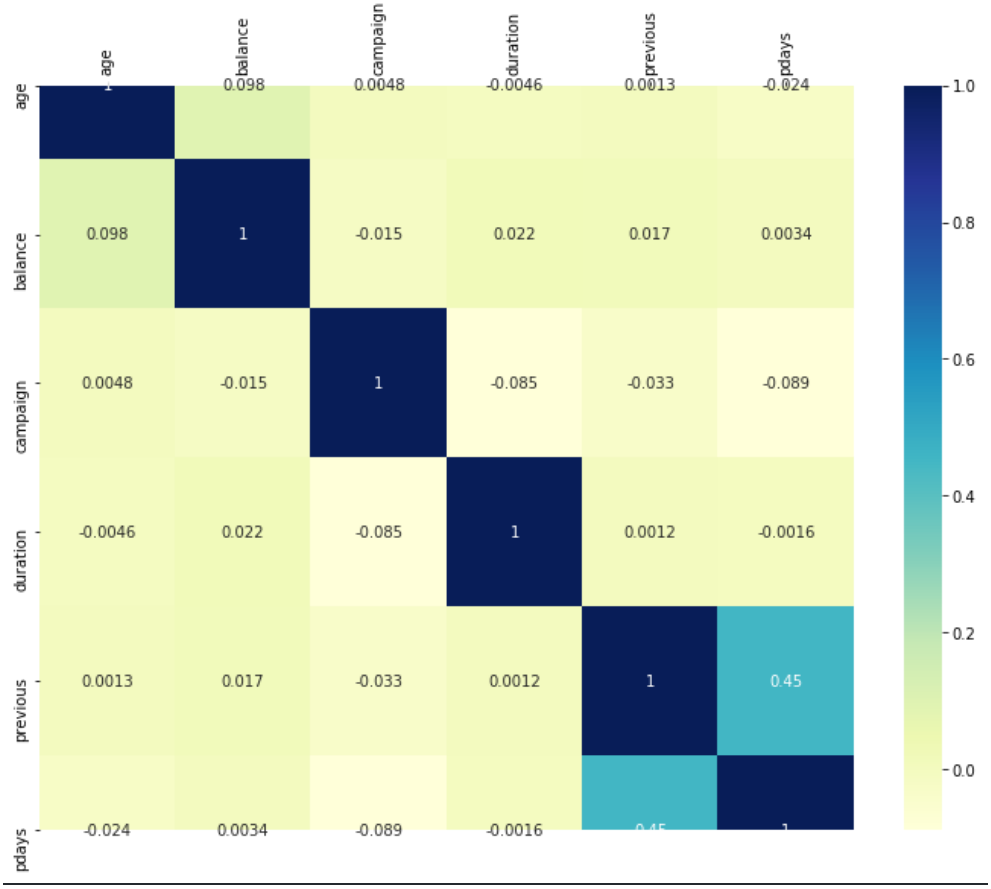
Histogram for Numeric Variable with respect to term deposit Flag = "Yes"



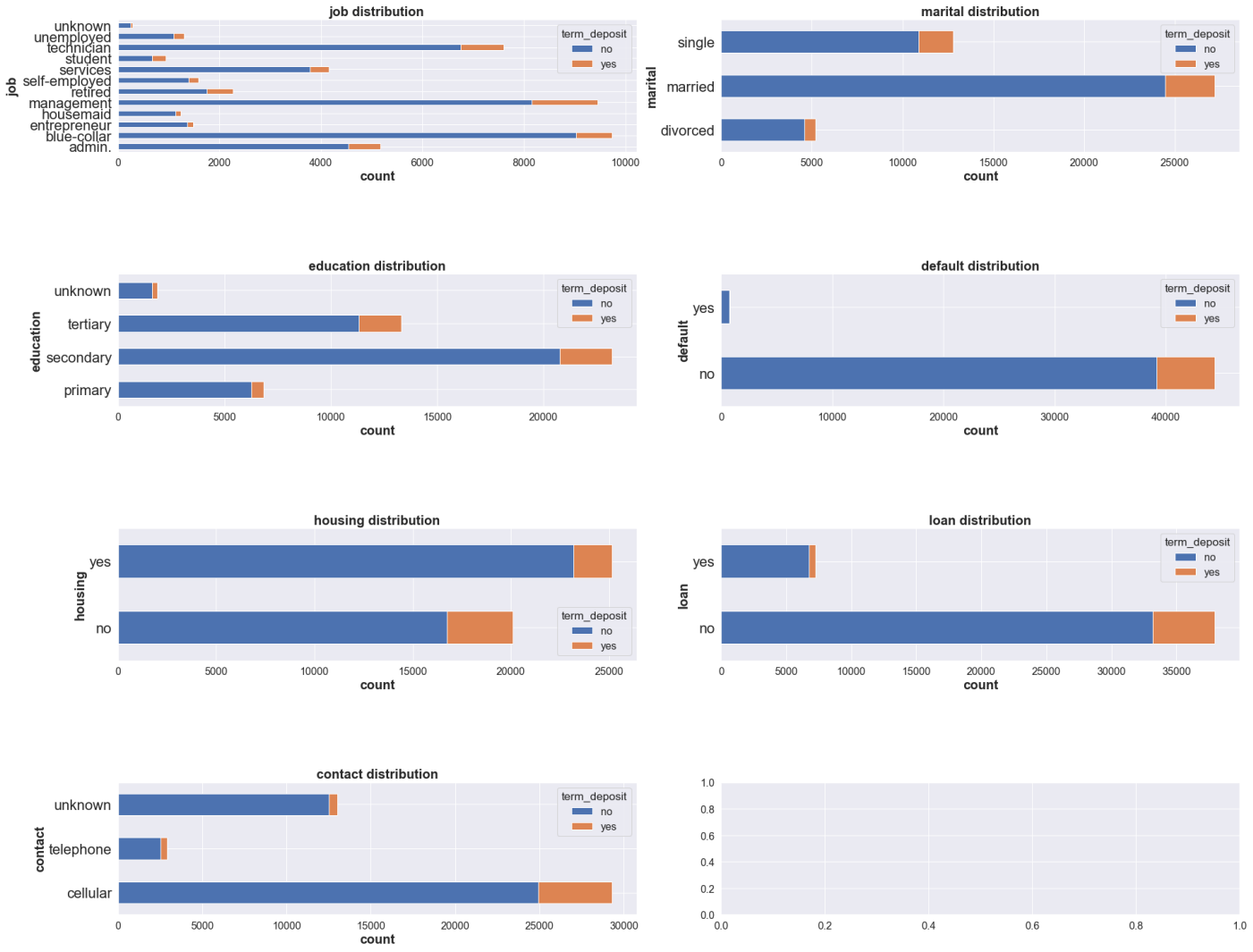
Observation:

- Middle age Customers have accepted the Term deposit offer
- Lower account Balance led to higher acceptance of term deposit offer
- Lower Contact call duration led to higher acceptance of term deposit offer
- Lower Campaign(# of contacts) call resulted higher acceptance of term deposit offer
- Low Previous # of contacts led to acceptance of term deposit offers

correlation heatmap of selected numerical features/predictor variables.



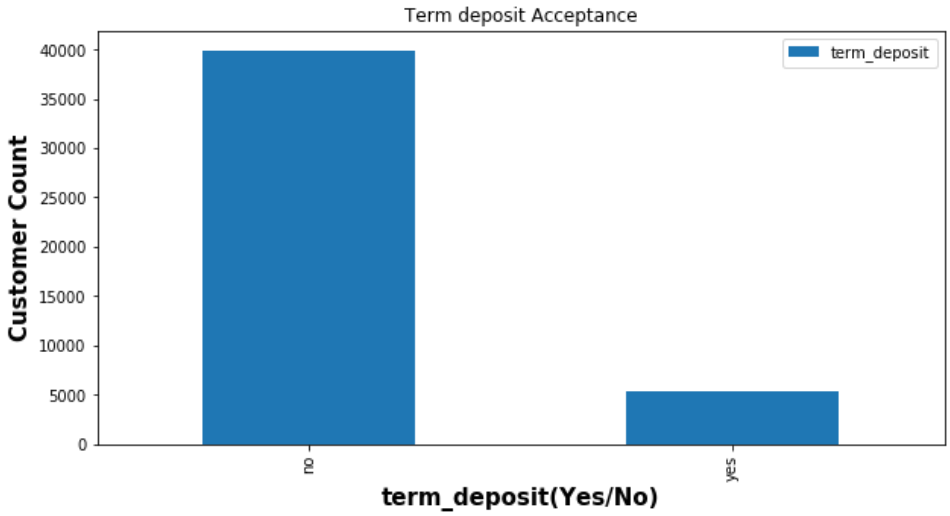
2. **For Categorical Variables**:
   - The Customers with following attributes have accepted the term deposit offers
     - Job (Management, Technician, Admin, Blue- Collar)
     - Married
     - Default-No Defaulted Customers
     - Education-Customers with tertiary/secondary education
     - Housing-Customers with (no Housing)
     - Loan-Customers with no loan
     - Contact-Customer contacted by Cellular



---

3. For Target (Response) variable-

- Term deposit (Yes/No)
  - Observation: **Imbalanced** Data set. As you can see, we have more term deposit (No) data points than term deposit(Yes) Data points



Algorithms and Techniques

We will be using Supervised Learning Binary Classification Model, i.e. Logistic regression to predict the Customer's acceptance of term deposit.

Justification: Logistic regression extends the ideas of linear regression to the situation where the outcome variable, *Y* , is **categorical**. In our project , we need to solve a binary Classification problem, which can be solved using Logistic Regression Model.

It uses logs of odds as Target/Dependent variable. it predicts the probability of occurrence of a binary event utilizing a logit function.

Additional Algorithm details:
The dependent variable in logistic regression follows Bernoulli Distribution. Linear regression is estimated using Ordinary Least Squares (OLS) while logistic regression is estimated using Maximum Likelihood Estimation (MLE) approach.

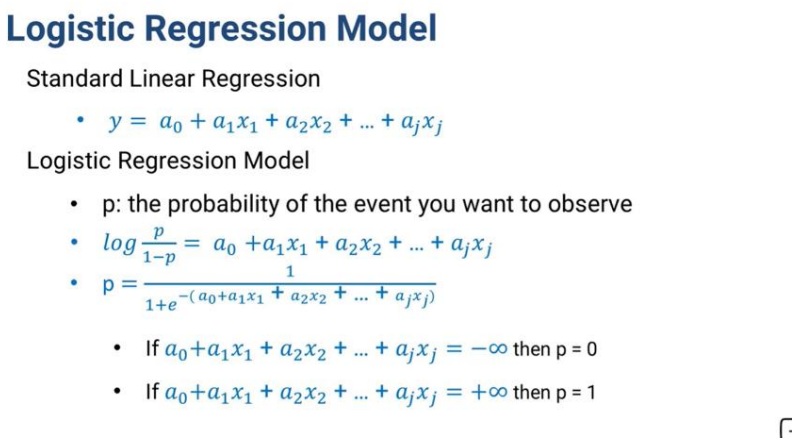
Mathematical presentation:
Let's say: Linear regression : y = β0 +β1X1 + β3X3+….βnXn
Where y = Target variable, X1, X2.. are predictor variables/features

Logit/sigmoid function: G(y) **= 1/(1 +e ^ -y)**

Since the labels are 0 or 1, we could look for a way to interpret labels as *probabilities* rather than as hard (0 or 1) labels. This is where we use the *logistic function*, also referred to as the ***logit* or sigmoid function**.

By applying the Sigmoid function to Linear Regression

G(y) = 1/(1 +e ^ - (*β0 +β1X1 + β3X3+….βnXn*)



The logistic function takes any value in the range (−∞,+∞)(−∞,+∞) and produces a value in the range (0,1)(0,1). Thus, given a value *y*, we can interpret *G*(*y*) as a conditional probability that the label is 1.

In summary: In logistic regression, we take primarily two steps:
(1) Derive estimates of the *propensities* or *probabilities* of belonging to each class. In the binary case, we get an estimate of $p = P(Y = 1)$, the probability of belonging to class 1 (which also tells us the probability of belonging to class 0).
(2) Then we use a cutoff value on these probabilities in order to classify each case into one of the classes.

Few Examples of application of Logistic regression is used in applications such as

1. Classifying customers as returning or non-returning (classification)
2. Finding factors that differentiate between male and female top executives (profiling)
3. Predicting the approval or disapproval of a loan based on information such as credit scores (classification)
4. Predicting the acceptance of fixed deposit offer based on marketing campaign (classification)

Below are the key steps that are followed in this project:

- **Pre-processing Data**
  - Detailed are outlined in next section

- **Splitting the data into Training and Test Sets**
    - We will use 80/20 ratio- 80% train Data, 20 % test data

- **Train the Model**

- **Evaluate the Model's effectiveness**
- Use Grid Search Cross validation technique to validate the effectiveness of the model and avoid the overfitting. The grid search package, **GridSearchCV**, uses cross validation to evaluate each model with different combinations of hyperparameters. This means that the data set is split into the number of folds specified (in this case, 10) and each unique model is tested with each fold of the data set and evaluated based on the specified metric (here, mean squared error).

- **Testing the Model**
    - Using 20% Test data

III. Methodology

Data Pre-processing

- Drop irrelevant predictor variables which have no influence on Customer acceptance of term deposit offer
- Drop rows with Nan/ missing values. *In fact, this data set has no Nan or missing values*
- Features Encoding (i.e. Convert Categorical data to numeric in order to train the model)
- Identify and remove outliers
    - Used z-score to remove outliers
- Scale and standardize the data as needed
- Create Final data set ready for Training/Test

Implementation

- **Pre-processing Data**
    - Detailed are outlined in above section

- **Splitting the data into Training and Test Sets**
    - We will use 80/20 ratio- 80% train Data, 20 % test data

- **Train the Model**

    - Using Logistic regression

- **Evaluate and Validate the Model**
  - Using GridSearch Cross validation

- **Testing the Model**
  - Using 20% Test data
  - Check the prediction accuracy using sklearn's metrics package. The details are outlined in <RESULTS> section

- Challenges/Difficulty faced during Developing the Model:
  - Also, there are few issues with Data Extraction step. Need to adjust the extraction step to pull the relevant features thru manual analysis and then performed the data clean-up/pre-processing as the part of Machine Learning workflow
  - Improving the accuracy- GridSearch CV didn't yield better accuracy after optimizing the various hyper parameters. The overall accuracy for Recall stayed at 81%, which is same as Base Model. I believe, I have used *Class Weight = "balanced"* in both Base and GridSearchCV Models. But the Best Score (i.e. mean cross validation score) was **87%** for GridSearchCV. So, I went ahead with GridSearchCV Model as the final model.

Refinement

I have used the **GridSearchCV** to choose the best hyperparameters combination to improve the accuracy.

I have used the following Hyperparameters for the GridSearchCV Model

```
Set hyperparamter options
penalty = ['l1', 'l2']
class_weight= ['balanced']
solver = ['liblinear', 'saga']
# Create regularization hyperparameter space
C = np.logspace(0, 4, 10)

# Create hyperparameter options
param_grid= dict(C=C, penalty=penalty,class_weight=class_weight,solver=solver)


grid_mod = GridSearchCV(logistic,param_grid=param_grid,cv=10,n_jobs=-1, verbose
=1,scoring='roc_auc')
```

As we can see, as the result of these model parameter changes, best_score went up to 87%.

```
best_score(Mean cross_validated score) : 0.8712323773325226
```

IV. Results

Model Evaluation and Validation

I have used the following Strategy for Model selection process

- Training Data set to fit model
- Validation Data set to choose best model (Model's **effectiveness)**
- Test Data Set to estimate **performance/quality** /**Evaluate** of Chosen Model

(1) GridSearch Cross validation- I have used the K-fold Cross validation to measure the effectiveness of the model. This step randomly divides the set of observations into k groups. The 1st set is treated as Validation set and the method is fit (Training) on remaining k-1 fold

```
Fitting 10 folds for each of 40 candidates, totalling 400 fits
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 8 concurrent workers.
[Parallel(n_jobs=-1)]: Done  34 tasks      | elapsed:   10.4s
[Parallel(n_jobs=-1)]: Done 184 tasks      | elapsed:  2.7min
[Parallel(n_jobs=-1)]: Done 400 out of 400 | elapsed:  3.5min finished
best_score(Mean cross_validated score) : 0.8712323773325226
Best Penalty: l1
Best C: 1.0
```

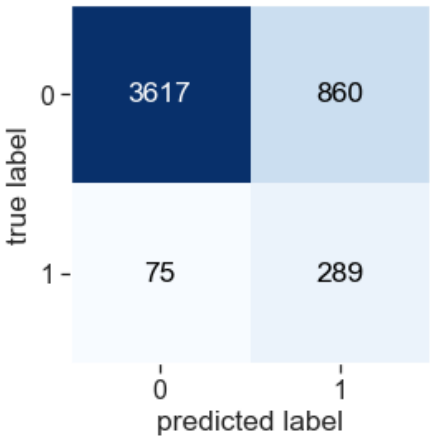
(1) Test the Model using Test Data

```
grid_mod = GridSearchCV(logistic,param_grid=param_grid,cv=10,n_jobs=-1, verbose
=1,scoring='roc_auc')
```

```
              precision    recall  f1-score   support

         0.0       0.98      0.81      0.89      4477
         1.0       0.25      0.79      0.38       364

    accuracy                           0.81      4841
   macro avg       0.62      0.80      0.63      4841
weighted avg       0.92      0.81      0.85      4841

Accuracy-FINAL Model: 0.8068580871720719
```

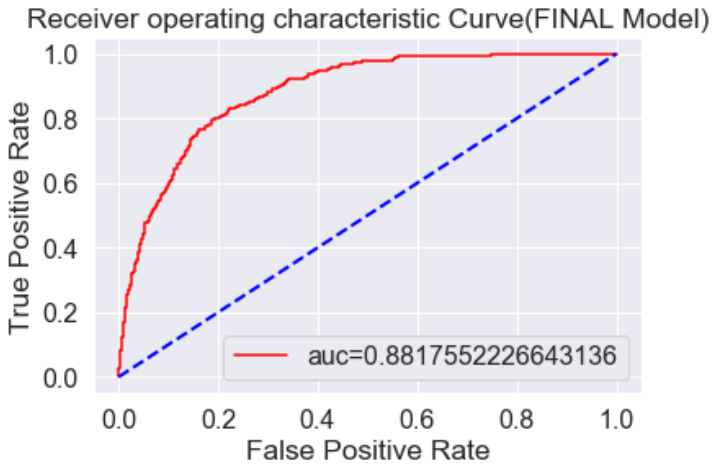
(2) Confusion Matrix

a. Model has correctly predicted 3617 customers to not accept the Term deposit offer(True Negative)
b. Model has correctly predicted 289 customers to accept the offer(True Positive)
c. Model has in correctly predicted 75 customers to accept the offer(False Positive)
d. Model has correctly predicted 860 customers NOT to accept the offer(False negative)



(3) Model -**ROC Curve**

Receiver Operating Characteristic(ROC) curve is a plot of the true positive rate against the false positive rate. It shows the tradeoff between sensitivity and specificity. The AUC score was **88%**

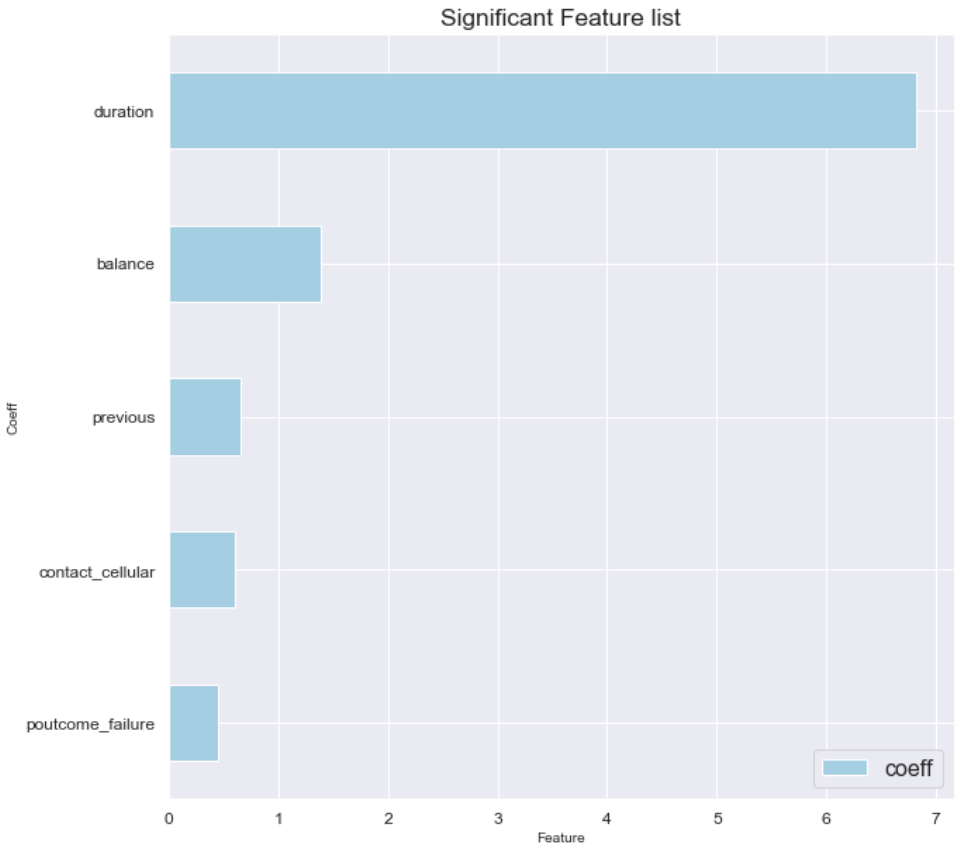


Based on Model effectiveness of 87**%** and Recall of 81**%**, I would believe our model is predicting the customer acceptance of term deposit offers correctly. However, the Model needs to get

updated/trained with new Ground Truth Labels (i.e. new Term deposit accepted data) to sustain the prediction accuracy.

(4) Significant Features from the Model

The following features/predictor variables are significant to drive higher response rate for term deposit offers

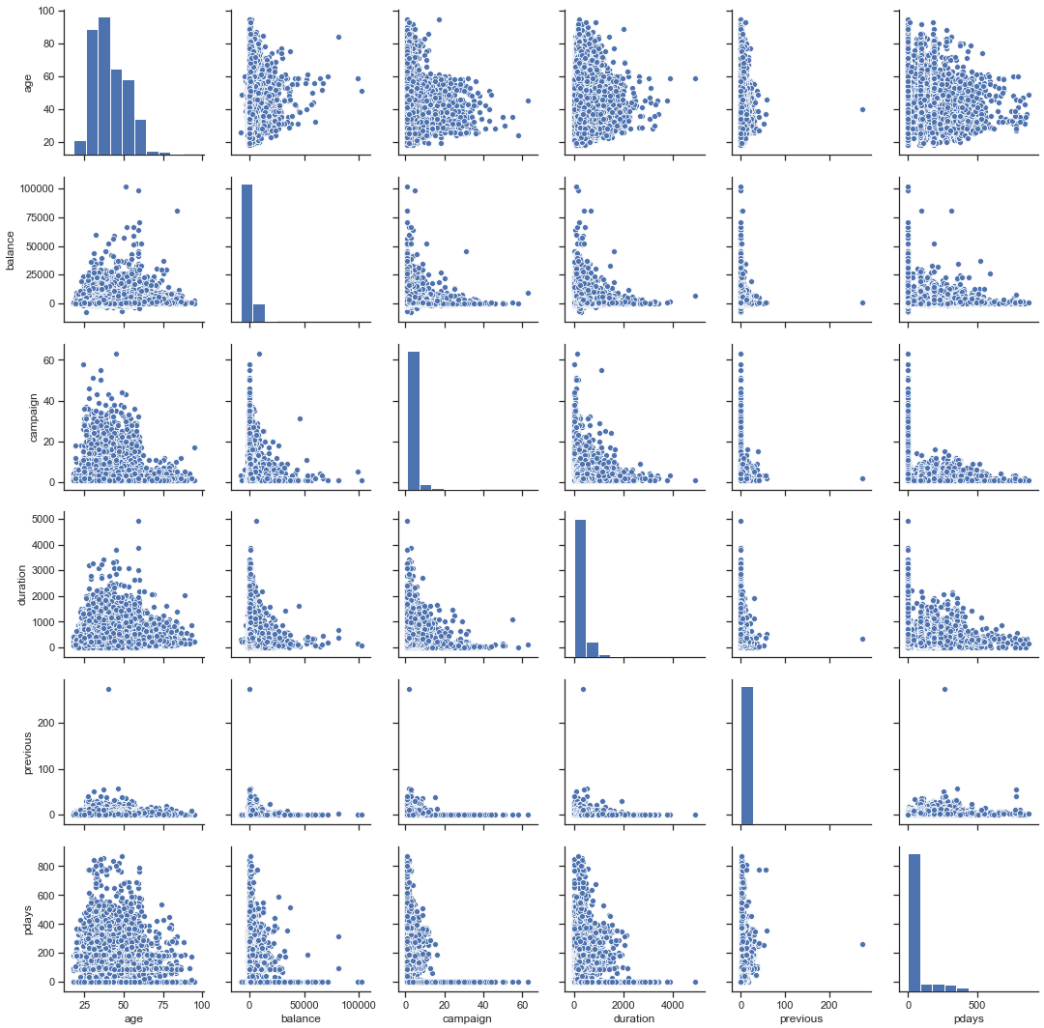


V. Conclusion

Free-Form Visualization

a. Visualizing pairwise relationship among features in a Dataset using Seaborn

- *The pairs plot builds on two basic figures, the histogram and the scatter plot. The histogram on the diagonal allows us to see the distribution of a single variable, while the scatter plots on the upper and lower triangles show the relationship between two variables*



Reflection

- I believe I have followed the Machine learning flows in this capstone Project, by leveraging Udacity excellent learning materials and 6 credit hours course work from Georgia Tech-Online MS in Analytics program.
- Still I think Modeling is an art. We get better at it as we use it more.
- Based on prediction recall (=81%), I think this model should it be used in a general setting to predict term deposit offers for Bank Customers.

Improvement

- **Model Tuning**
    - Need to retrain the model with ground truth (i.e. new Bank customers data with accepted term deposit offers) to boost model prediction
    - Need to leverage Amazon Sage Maker or any other Cloud Machine learning Platform to avail Model tuning options.
    - We can use sklearn 's RandomizedSearchCV to perform random search on Hyper Parameters
    - We could have used different classification models such as:
        - Gaussian Naive Bayes
        - Decision Tree
        - Random Forest
        - XGBoost

- **Training Set Quality**
    - Need to resample data in order to get a "balanced" data set
    - Need to ensure all the required predictor variable data were populated
    - Instead of removing rows with Missing data, need to assess the data imputation methods (such as Mean/Mode/Median, Prediction model etc.)

References

1. Choosing the right metric Choosing the Right Metric for Evaluating Machine Learning Models.
2. Article about What metrics should be used for evaluating a model on an imbalanced data set
3. Article on various techniques of the data exploration process.
4. Data Mining for Business Analytics- *Galit Shmueli, pertr C Bruce ,Wiley Publications*
5. DataCamp: Logistic regression :https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python