

Tb

**TABLE OF CONTENTS**

[**Executive Summary**](#_x8664hoqre4q) **3**

[**Statement of Scope**](#_g7b2w4vk24gl) **4**

[**Project Objectives**](#_gnphsov680io) **4**

[Predictor Variables](#_kzr67w7ttivy) 5

[Target Variables](#_2cj59spwaftd) 5

[**Project Schedule**](#_wjh763lnyhc1) **6**

[**Data Preparation**](#_ywiki5uceqau) **7**

[Data Access](#_p4ylo76vlfc0) 7

[Data Consolidation](#_z2n2k3tod4ey) 7

[Data Cleaning](#_ceh6uwvgyyq7) 9

[Data Transformation](#_yl0m3kw179f5) 14

[Data Reduction](#_jqc9g3mvunr7) 15

[Descriptive Statistics](#_rawsv06o5aa1) 17

[Data Dictionary](#_amt5aul19idl)

**Modeling Techniques** 27

Decision Tree

Logistic Regression

**Model Assumption** 30

Decision Tree

Logistic Regression

**Data Splitting and Subsampling** 30

Comparison of variables between test and training data

**Model Building** 33

**Model Assessments** 38

**References** 39

## Executive Summary

We are working on the Teradata 2019 Data Challenge where we are provided with multiple data sets and business questions from Hire Heroes USA: one of the non-profit partners of Teradata. Hire Heroes USA provides personalized job-search assistance to U.S military, veteran and military spouse community. In particular, its clients are aided in resume review, workshops, webinars, networking, mock- interviews, job-search etc.  
 For this data challenge, we have analyzed the data related to the ‘Hire Heroes USA: Clients Services program’ to provide valuable insights into how the clients demographic profile influences when a particular client registers for services and the likelihood that a client completes a survey. These results helps Hire Heroes USA to understand the patterns and behaviour of its clients. Additionally, the results will help Hire Heroes USA to operate effectively and increases the usability of their system and help in better transition of its clients.

## Statement of Scope

The dataset contains data of 1,32,430 records comprised of US military members, veterans and military spouses. It contains the clients data such as branch, pay grade, service time, spouse status etc. The following business questions will help in deriving insights about how to improve the enrollment process of Hire Heroes. The analysis can potentially save clients time and can significantly smoothen the process of hiring. This analysis also helps veteran service organizations and federal agencies in better understanding of its clients (military, veteran and military spouse community).

* The relationship between a client's demographic profile (rank, branch, time in service etc.) and when that client registers for services?
* How are factors like rank, branch, spouse status etc. are affecting the client’s likelihood to complete a survey?

### Project Objectives

* By getting insights about the likelihood to complete a survey; different methods can be established that can help gathering information by Hire Heroes.
* The process of hiring can be improved to a great extent when hidden information about the behaviour of the members would be unveiled. Changes can be implemented accordingly by knowing which ranks, branches registers for the services earlier or later that helps in preparing in advance and help queuing tasks.

**Business question 1:** The relationship between a client's demographic profile (rank, branch, time in service etc.) and when that client registers for services?  
Target Variables:

* Joining\_Period : which is Difference between military separation date and created date in Hire heroes USA

### Predictor Variables

* Col : which is Administrative classification of ranks by pay grade and branch of military service
* Service\_Branch\_\_C : which is the Branch of military service (Army, Navy, etc.)
* Service\_Period : which is the Time in military service in months (difference between separation date and entry date in military)
* Military\_Spouse\_Caregiver\_\_C : which is True / False (if true, indicates job seeking client is a spouse of veteran / servicemember and does not have military service)

**Business question 2**: How are factors like rank, branch, spouse status etc. are affecting the client’s likelihood to complete a survey?

### Target Variables

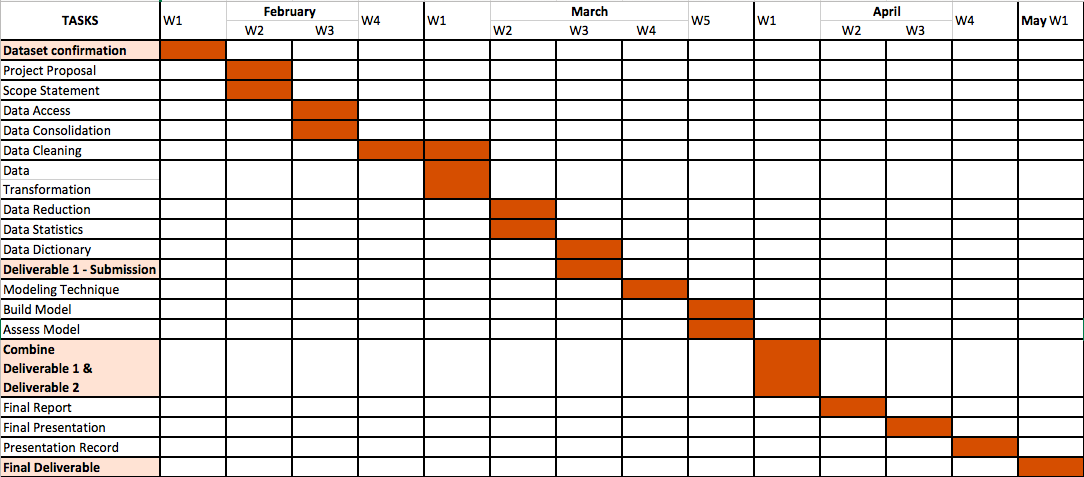
* Alumni\_Survey\_Completed\_\_c: True / False (if true, client responded to / completed alumni program survey)

**Predictor Variables**: (same as for business question 1)

* Col : which is Administrative classification of ranks by pay grade and branch of military service
* Service\_Branch\_\_C : which is the Branch of military service (Army, Navy, etc.)
* Service\_Period : which is the Time in military service in months (difference between separation date and entry date in military)
* Military\_Spouse\_Caregiver\_\_C : which is True / False (if true, indicates job seeking client is a spouse of veteran / servicemember and does not have military service)

## Project Schedule

The team anticipates to complete the project in 89 days. The project meet is scheduled two times in a week. We generally discuss what each member worked on the week before and what we should work on the current week. Individually, we need 4 to 5 hours to complete the weekly work. We don’t work on weekends. If there is a holiday the day we were supposed to meet, we generally meet the day later and never delay on our weekly assignment.



## Data Preparation

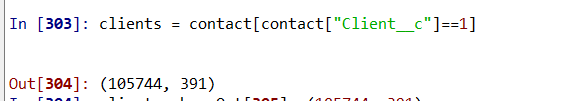
### Data Access

The datasets were obtained through Teradata 2019 Data Challenge. The files are of type .csv There are 13 datasets given as a part of the Challenge which sum to 764 MB. We are using only 1 dataset of 209 MB which contains all the information required for Client Services. We also required the grouping category name i.e military rank ( Warrant Officer, Enlisted Members, Commissioned Officers) for different Branch ( Army , Navy, Marines etc.) based on their pay-scale. The data for this was not available in the teradata data sets. The required military rank was available at <https://militarybenefits.info/armed-forces-comparative-pay-grades-and-ranks/> So, we imported the data available at the above link into a csv from <https://app.import.io/> website. The import process was an easy one. We started with what specific columns and rows we required and clicked on import and then downloaded.This was a 5 KB file. We later consolidated the columns in a new file along with the required columns for our analysis.

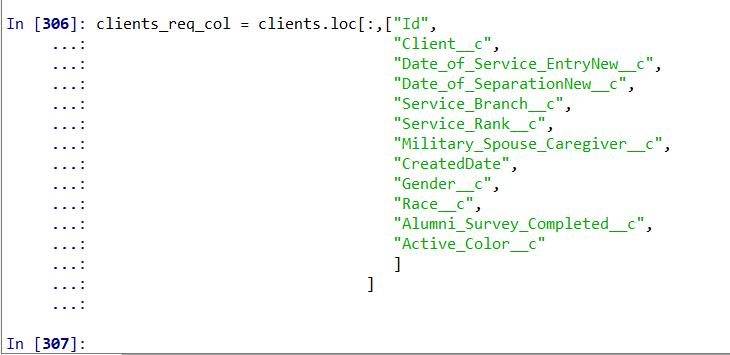
### Data Consolidation

We used SalesForce\_Contact.csv and MilitaryRank.csv files to consolidate the data.

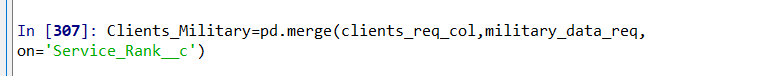
SalesForce\_Contact.csv file contains the data of all the people who are associated with HireHerosUSA (Ex: Clients/Jobseekers, Volunteers, Donors, Transition Specialists). As our analysis is completely based on Clients/Jobseekers this data has been filtered based on Client\_\_c=1. So this brings down the data to 105744 rows.



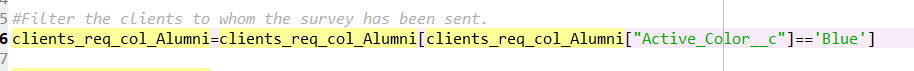
All the relevant columns that will help us define rank, branch, time in service, spouse status, when client registered and likelihood to complete Alumni Survey are selected from this data set.

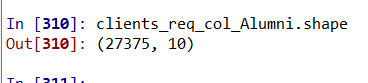


Service\_Rank column has the codes based on the levels of latest pay grade of the clients when they register at HireHerosUSA. To make it more readable by the business we have used external data source MilitaryRank.csv. Both the data sets have been merged based on the Ranks.



After this point we need two different data sets to answer these questions. Because the first question talks about when the client joins HireHerosUSA, in this case we can consider all the clients who are part of the system. But, for the second question we are talking about Alumni Survey which will be applicable to the clients who are already hired. i.e with Color\_Code = Blue.

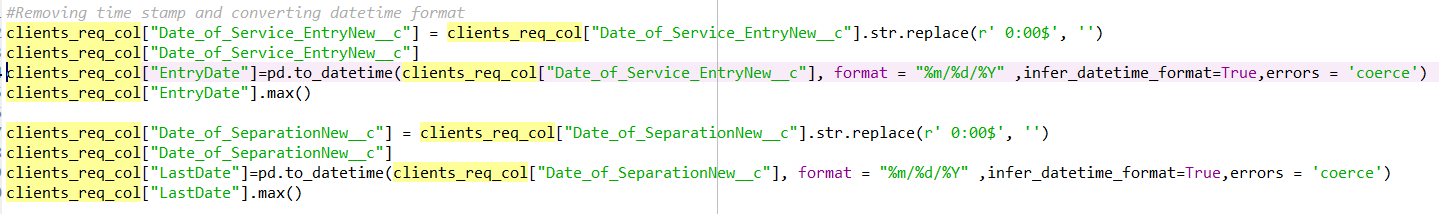


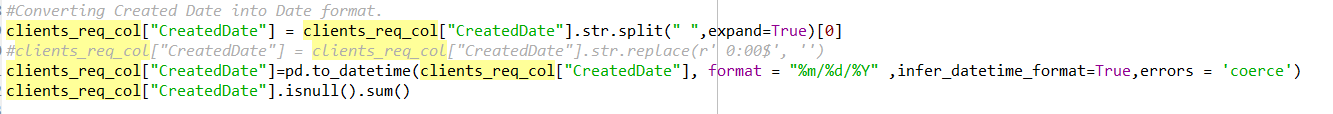


### Data Cleaning

**Converting Dates**

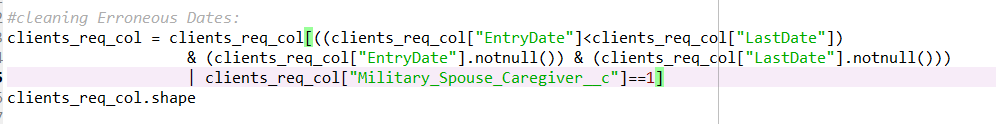
We have three Date Columns in our dataset(Service\_Start\_Date, Service\_End\_Date and Created Date). All these variables are in the object form. along with TimeStamp. But while calculating Service\_Period (Time in service) and Joining\_Period (When client registers with HireHerosUSA) we dont require the time stamp. So the time stamp portion of the Date string has been trimmed with the help of Pandas String functions (replace and split)





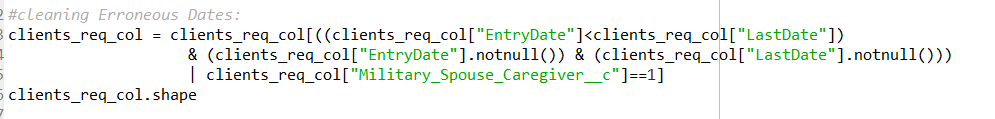
**Deleting Erroneous Data:**

When we are calculating the variable Service\_Period (Time in Service) we saw some negative values. Which shows that the data has some incorrect entries where Service\_Start\_Date is later to the Service\_End\_Date. So this data has been eliminated deleted from the data set by checking for the rows with Start date later than End Date while doing this we need to make sure that we don't lose the spouse data as the date column may have null values in these columns.

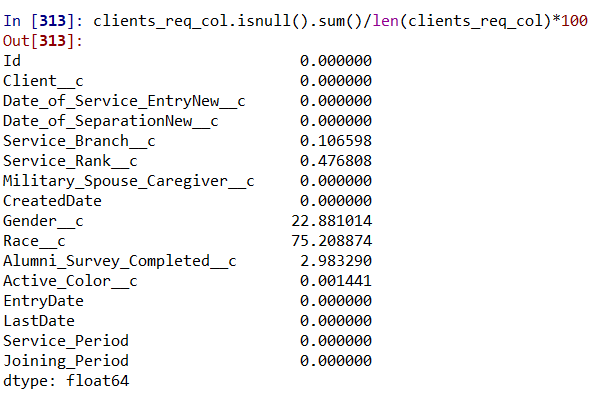


**Deleting the NaT values of Dates:**

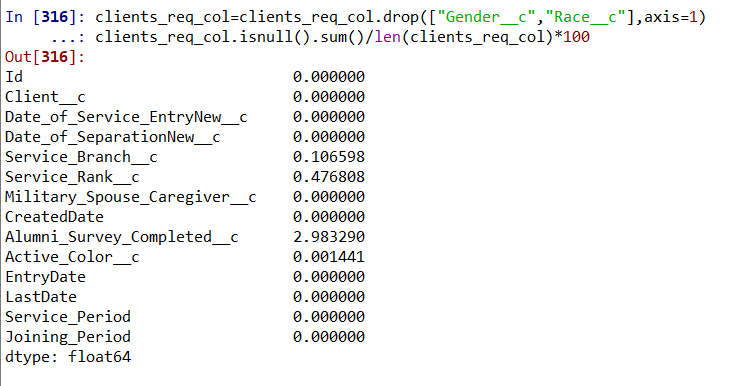
While transforming the Date Columns to DateTime the function pandas.to\_datetime will return NaT values when the Date is greater than 2270-12-31. This would create Na values all the there Date columns after the transformation. These rows should be filtered out.



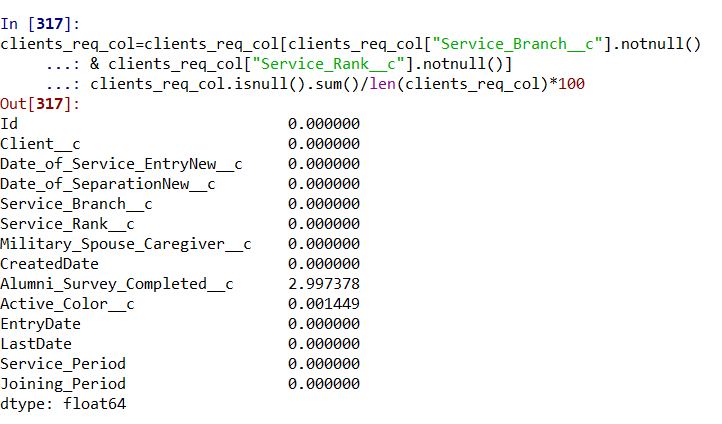
**Checking Na values of other columns.**



From the above table we can see that Gender and Race columns have more than 20% of the data so we can completely delete these columns from the Data.



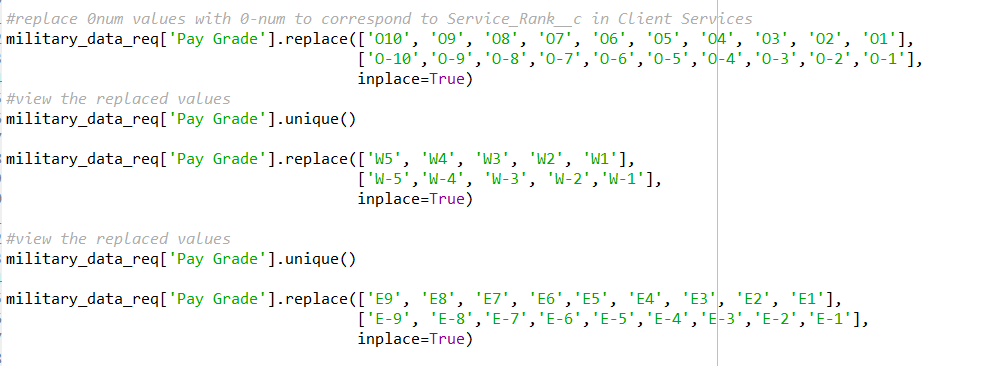
The rest of the variables have missing values of less than 3% so, we can delete the rows corresponding to these missing values.



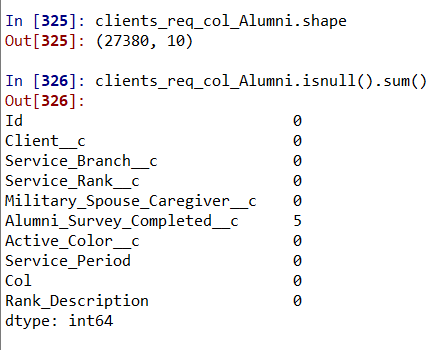
From the updated table we can see that only Alumni\_Survey\_Completed\_\_c and Active\_Color\_c variables have missing values. But these columns are required only to answer second question (Likelyhood of the survey) So by removing these columns we create the final data set required for the analysis of relation between the demographics and when the client has joined HireHerosUSA.

**Aligning the Rank values in both data sources to merge**

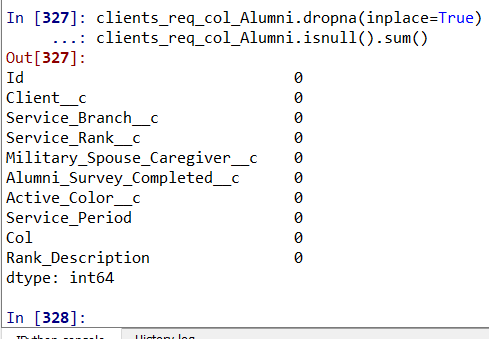
The Rank values have been adjusted in the external data source in order get them aligned with the values in HireHerosUSA datasets.



**Removing Null values for data related to Question 2**

****

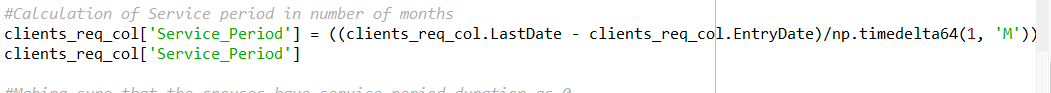
From the table we can see that just 5 rows has na values. These rows have been eliminated from the dataset in order to get the final data set required for the analysis of likelihood of survey completion.



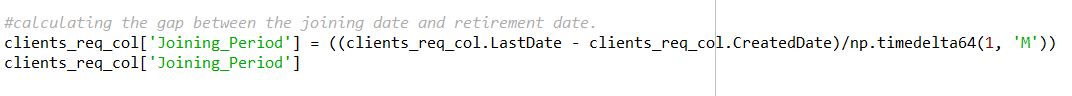
### Data Transformation

**Creation of Calculated variables**

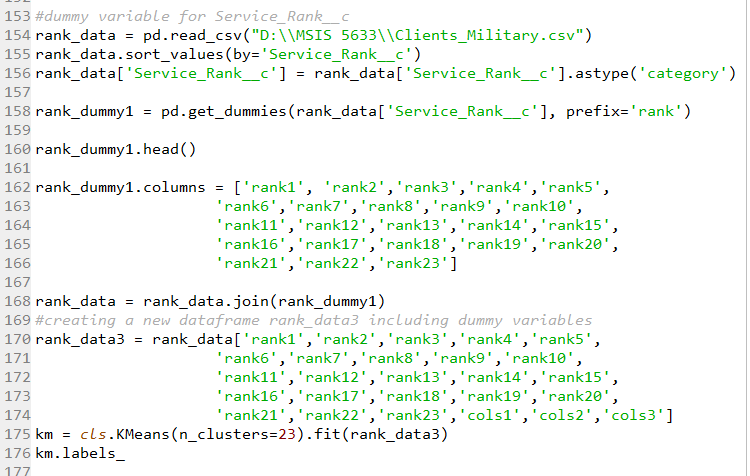
From the StartDate and EndDate we derived the duration of clients service at armed forces in number of months.

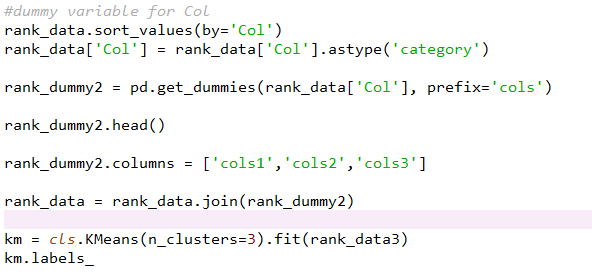


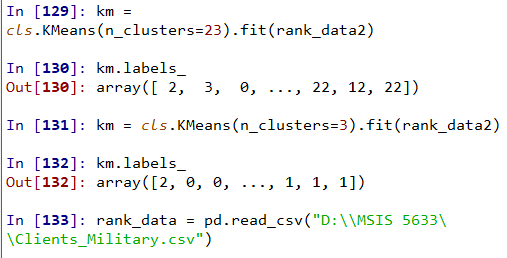
From the EndDate and Created Date we derived the Joining period (How early or late is he joining HireHerosUSA from his last day at armed force in months)



### Data Reduction

****

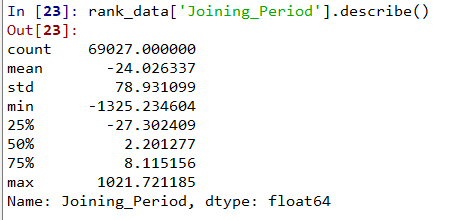
****

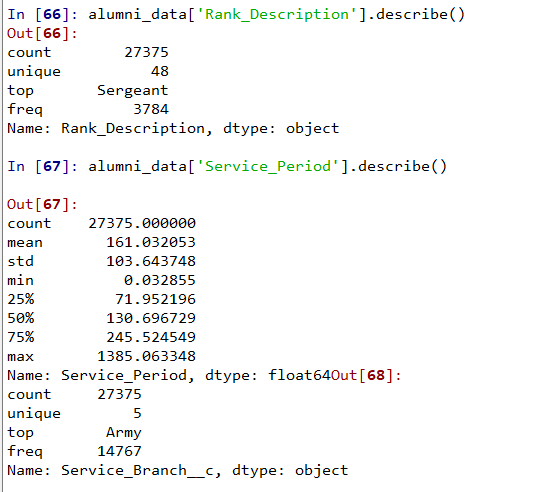


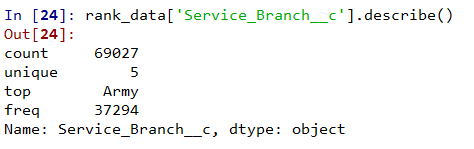
We have the Ranks classified into 23 categories from the HireHerosUSA data with various pay level codes and we also have high level classification these Ranks into three categories. As both the variables represents the pay level it is good to get rid of one of these two. To know which variable best explains the variation in data we should apply one of the Data Reduction Techniques. As we are looking at Categorical variables it is best to use Cluster analysis as data reduction technique in this case.

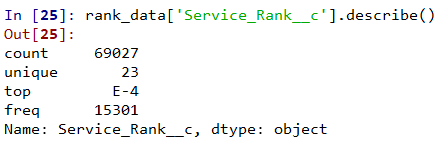
Above are the results of the K-means cluster analysis that has been performed on the data set with 23 clusters and 3 clusters based on the number of unique values in each of the Rank related variables. From the results we can see that the variable ‘Col’ (3 categories) best explains the classification of data in our data set. As a result we will eliminate the Service\_Rank\_\_c column from the data set.

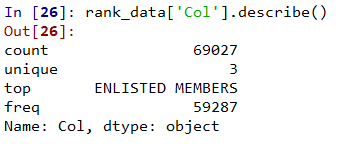
### Descriptive Statistics

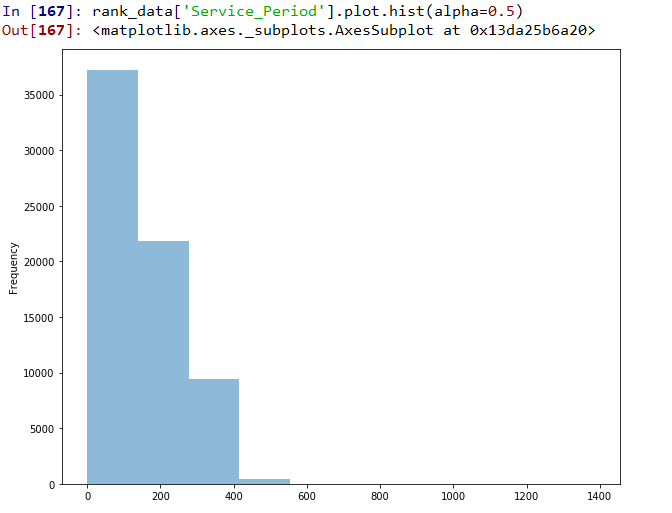
The descriptive statistics below shows the mean shows the means, standard deviation, Minimum and maximum values for all the variables 

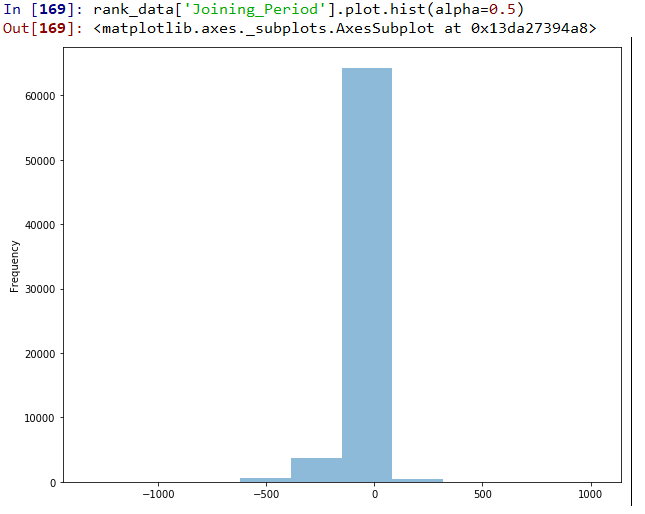


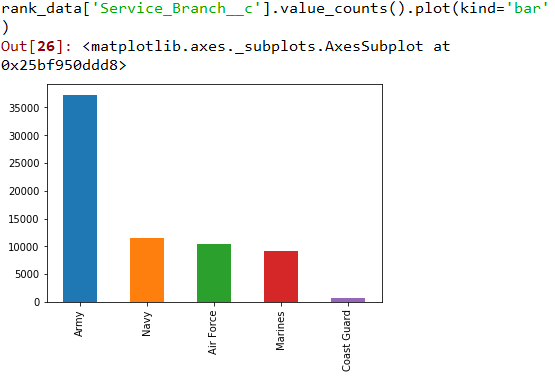


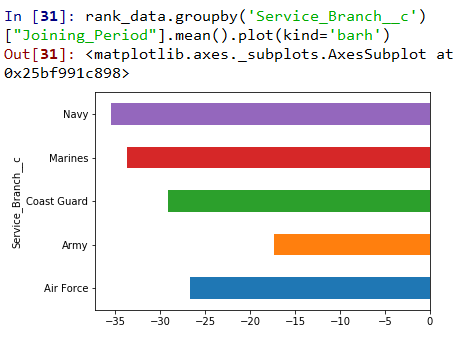


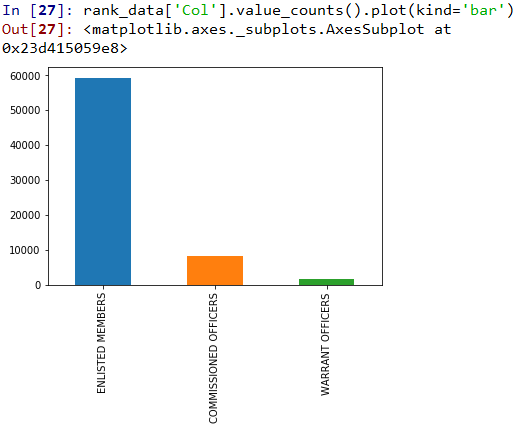


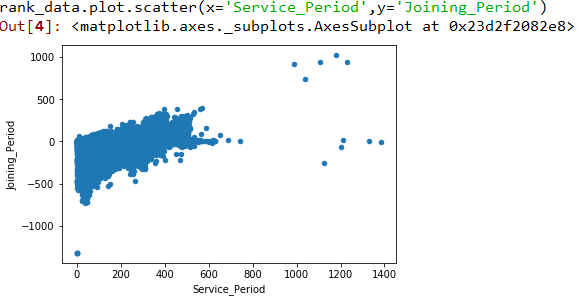




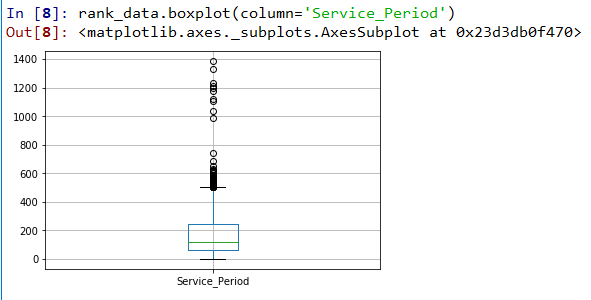


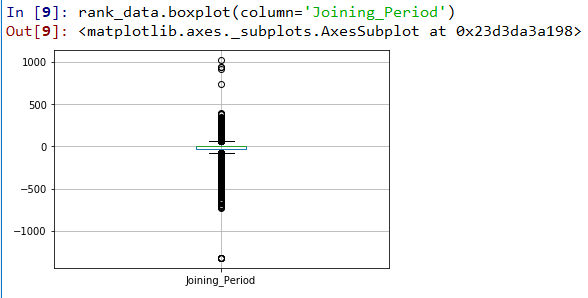
Graph shows the count of the members according to their branches.



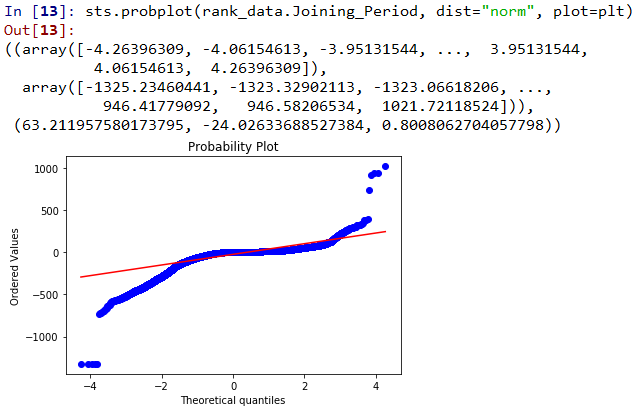


The above graph shows somewhat linear pattern in the initial ranges and we can clearly observe few outliers.





We can observe outliers in the above Box Plots.



The data is somewhat heavily tailed.

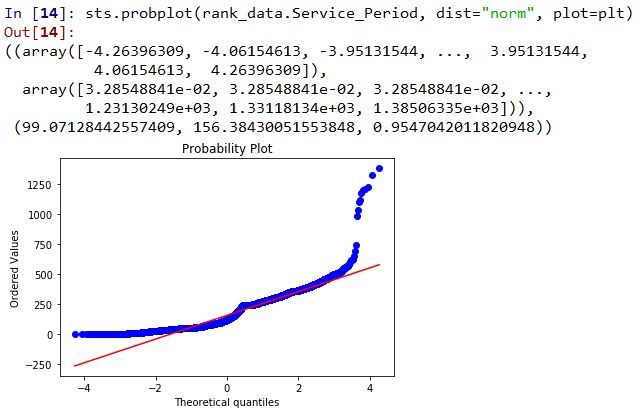
sts.shapiro(rank\_data.Joining\_Period)

Out[64]: (0.6414784789085388, 0.0)  
Note: p-value may not be accurate for N > 5000.

In the above Shapiro-Wilk test output,

first argument = value of test statistic and second argument = p-value. We find p-value is less

than 0.05. Hence we reject the null-hypothesis that radiation is normally distributed.



Data is observed in bow shape and by which we can say it is definitely not linear.

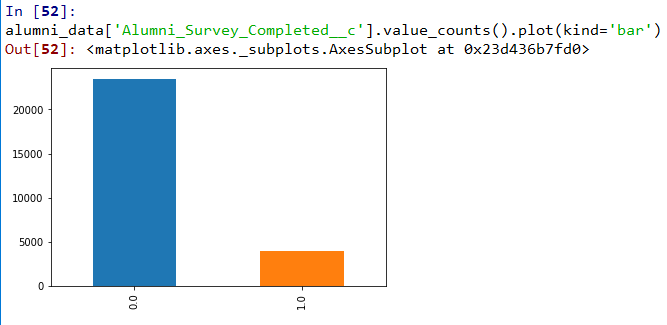
sts.shapiro(rank\_data.Service\_Period)

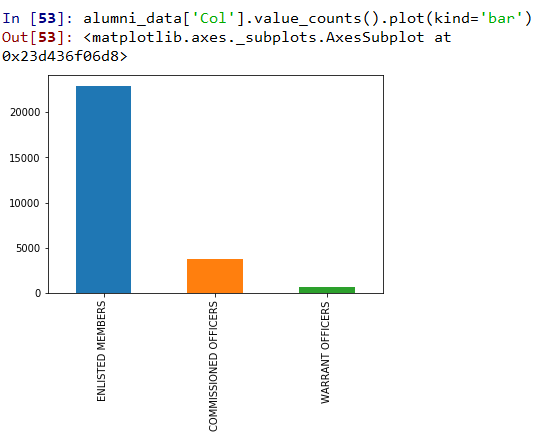
Out[65]: (0.9114835858345032, 0.0)

From the above Shapiro-Wilk test output, We find p-value is less

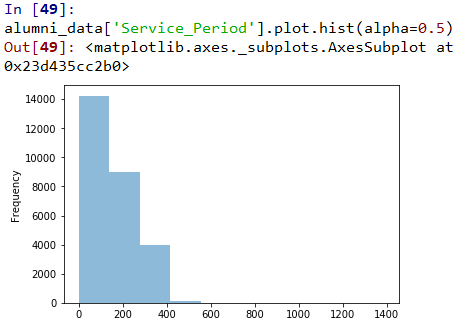
than 0.05. Hence we reject the null-hypothesis that radiation is normally distributed

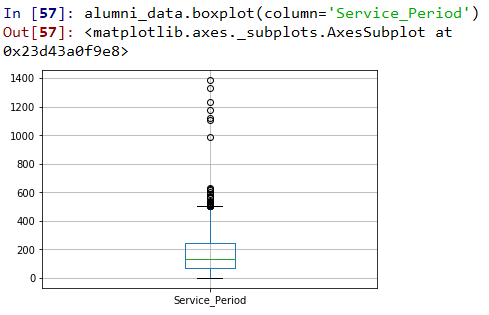
In the below plot, we see that the ones who complete the survey are much less than the ones who complete.



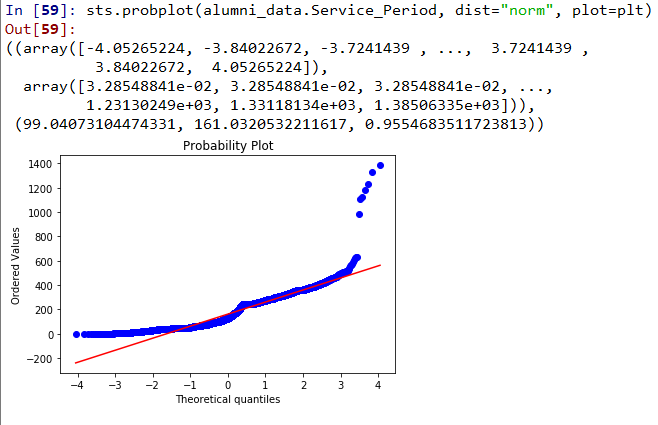


The following graph shows the member’s categories and the count having Enlisted Members at the highest followed by Commissioned Officer and Warrant officers with the least value.





From the boxplot we can clearly see outliers are at a large number.



The above plot show a bow shape pattern .Thus we can see that it is not linear.

sts.shapiro(alumni\_data.Service\_Period)  
Out[62]: (0.91303551197052, 0.0)

### Data Dictionary

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute Names** | **Description** | **Data types** | **Source** |
| Id | Unique Salesforce Identifier (individual contact) | Object | <https://www.teradatauniversitynetwork.com/Community/Student-Competitions/2019/2019-Data-Challenge/Data-Challenge-Datasets> |
| Client\_\_C | True / False (Indicates the account is that of a job seeker) | Integer | <https://www.teradatauniversitynetwork.com/Community/Student-Competitions/2019/2019-Data-Challenge/Data-Challenge-Datasets> |
| Date\_of\_Service\_EntryNew\_\_C | First day of client's military service | Object | <https://www.teradatauniversitynetwork.com/Community/Student-Competitions/2019/2019-Data-Challenge/Data-Challenge-Datasets> |
| Date\_of\_SeparationNew\_\_C | Last day of client's military service (actual or anticipated) | Object | <https://www.teradatauniversitynetwork.com/Community/Student-Competitions/2019/2019-Data-Challenge/Data-Challenge-Datasets> |
| Service\_Branch\_\_C | Branch of military service (Army, Navy, etc.) | Category | <https://www.teradatauniversitynetwork.com/Community/Student-Competitions/2019/2019-Data-Challenge/Data-Challenge-Datasets> |
| Service\_Rank\_\_C | Most recent pay grade of job seeking client (if military service / E-4, O-7, etc.) | Category | <https://www.teradatauniversitynetwork.com/Community/Student-Competitions/2019/2019-Data-Challenge/Data-Challenge-Datasets> |
| Military\_Spouse\_Caregiver\_\_C | True / False (if true, indicates job seeking client is a spouse of veteran / servicemember and does not have military service) | Integer | <https://www.teradatauniversitynetwork.com/Community/Student-Competitions/2019/2019-Data-Challenge/Data-Challenge-Datasets> |
| CreatedDate | Date unique account was established | Datetime | <https://www.teradatauniversitynetwork.com/Community/Student-Competitions/2019/2019-Data-Challenge/Data-Challenge-Datasets> |
| EntryDate | First day of client's military service | Datetime | Manual creation |
| LastDate | Last day of client's military service (actual or anticipated) | Datetime | Manual creation |
| Service\_Period | Time in military service in months (difference between separation date and entry date in miltary) | Float | Manual creation |
| Joining\_Period | Difference between military separation date and created date in Hireheroes | Float | Manual creation |
| Col | Administrative classification of ranks by paygrade and branch of military service | Category | <https://militarybenefits.info/armed-forces-comparative-pay-grades-and-ranks/> |
| Rank\_Description | Listing of ranks by paygrade and branch of military service | Category | <https://militarybenefits.info/armed-forces-comparative-pay-grades-and-ranks/> |
| Alumni\_Survey\_Completed\_\_c | True / False (if true, client responded to / completed alumni program survey) | Integer | <https://www.teradatauniversitynetwork.com/Community/Student-Competitions/2019/2019-Data-Challenge/Data-Challenge-Datasets> |

Adjusting the Deliverable 1 Requirements  
**General Project Information**  
We have not made any changes in the title of our project or the team members.  
**Executive Summary**

For this data challenge, we have analyzed the data related to the ‘Hire Heroes USA: Clients Services program’ to provide valuable insights into how the clients demographic profile influences when a particular client completes an alumni survey.  
**Project schedule and duration changes**We did not make any changes to our initial schedule. We have planned our schedule very well since the beginning.

**Statement of Scope**For the deliverable 2, we will be concentrating only on the 2nd business question which is : how the factors like rank, branch, spouse status etc. affect the client’s likelihood to complete a survey? . That is we are trying to understand whether the factors like the branch of military or the pay grades of the clients or their time in military service helps to determine whether an alumni (who has a job through the guidance of Hire Heroes) completed the alumni program survey or not. This will help in knowing the clients better and smoothen the process of surveys going forward by targeting specific areas concerned.

# 

# Modeling Techniques:

We are using the following two modelling techniques in our project.

## Decision tree

The model Decision Tree is chosen as the first model for analysis. This model generally uses divide and conquer method to represent data. This algorithm helps in visualizing the decision making processes wherein an optimal solution is presented at the end of each node. The interpretations are easily explainable which is one of the factors giving Decision Tree an edge over other algorithms.

## Logistic Regression

Logistic Regression was chosen as the second model for the analysis as the algorithm is apt for our case where the dependent variable is categorical. In our model, the target variable “Alumni\_Survey\_Completed\_\_c” is a binary categorical variable which has values of ‘0’ or ‘1’ .

# Model Assumption

## Decision tree

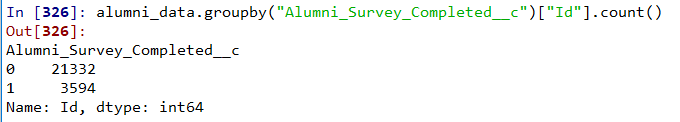
Being a non parametric algorithm there are not any assumptions made on the distribution of data or errors. In other words the type of data decides the construction of the model.

## Logistic Regression

Logistic Regression demands the variables to be independent of each other. The multicollinearity is assumed to be very little or not at all among the independent variables. The independent variables and log odds are assumed to be linear. Having a fairly large data set in one of the assumed conditions with the absence of any outliers in the continuous predictors. These assumptions have been met in the model.

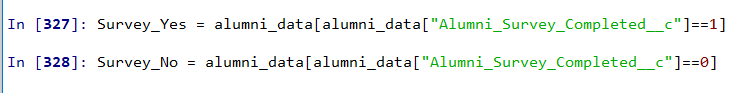
# Data Splitting and Subsampling

In order to estimate the likelihood of an Alumni Completing a Survey we need to run regression model with target variable as “Alumni\_Survey\_Completed\_\_c”. Below is the distribution of the target variable.

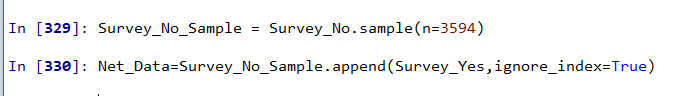


In order to continue with Logistic regression we need to ensure equal variance among the variables. One way to ensure equal variance is to do stratified sampling.

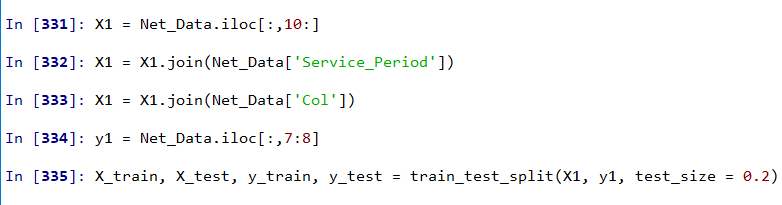
This has been achieved by splitting the data into two parts, one with Alumni\_Survey\_Completed\_\_c variable = 0 and the other with Alumni\_Survey\_Completed\_\_c =1.



And then the 0’s are randomly sampled to pick the same number of records as of 1’s. and the sampled 0’s are appended with the 1’s dataset.

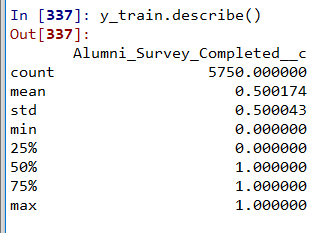


The resultant dataset is now split into testing and training data sets with 20% and 80% partitions respectively.

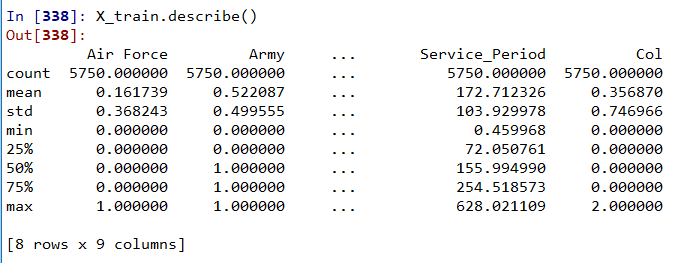


#### Comparison of variables between test and training data

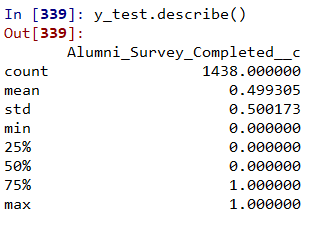
**Training Data: target variable**



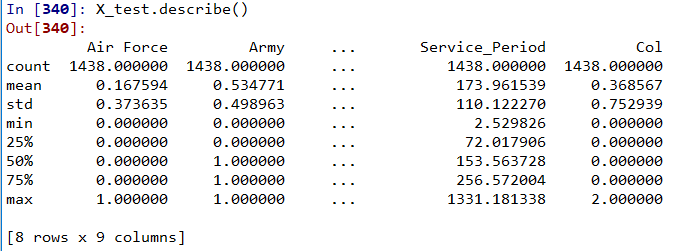
Training data input variables



**Testing Data: target variable**



describing the input variables of the testing data

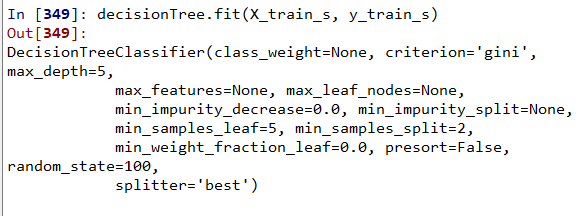


From the above results we can clearly see that, all the variables are exhibiting similar nature across both test and training datasets.

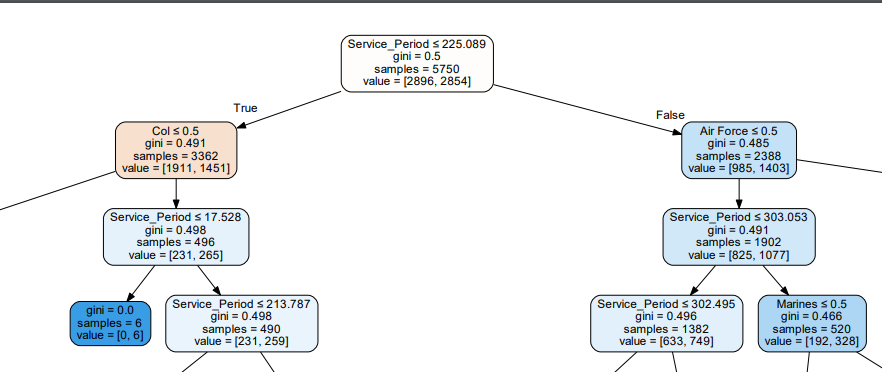
**Model Building**

## Decision Tree Model

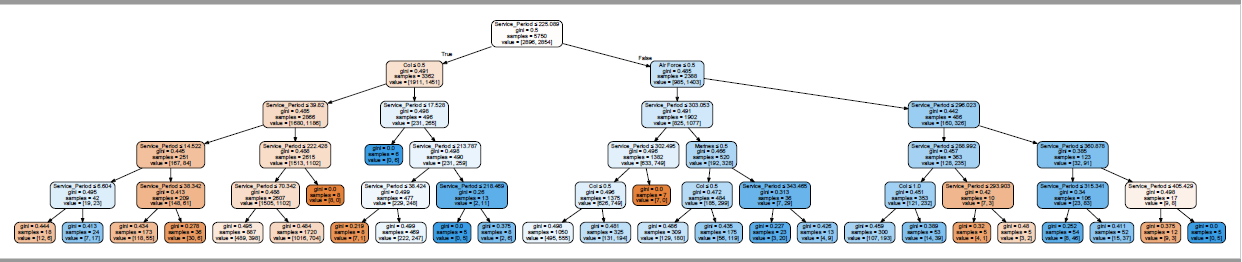
As the target variable is categorical variable we considered running Decision Tree Model. The model has been run with max\_depth of 5 levels. Below are the results of the model.



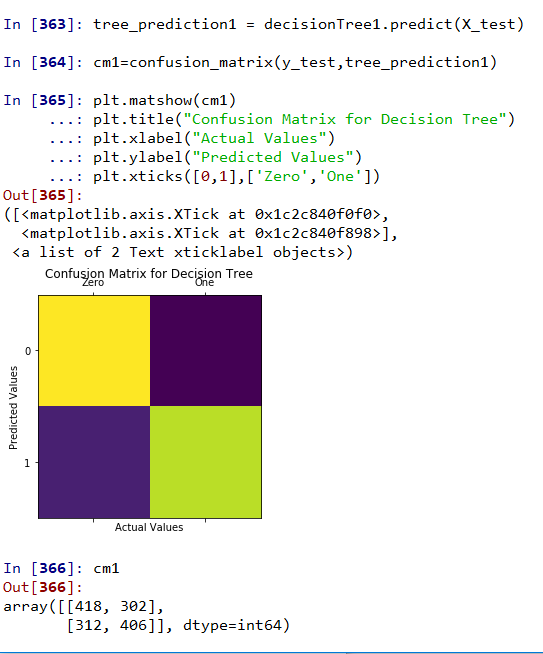
First level of decision tree



Based on the above tree, we see that the first criteria for the split is whether the service\_period is less than or equal to 225 months or greater than this. If the service period is greater than the one specified, the fact whether the military category is Air force less than or equal to 0.5 is the next criteria for split. Depending upon the service period, the payscale of the military or veterans increase.



**Classification Matrix:**



Results and Interpretation

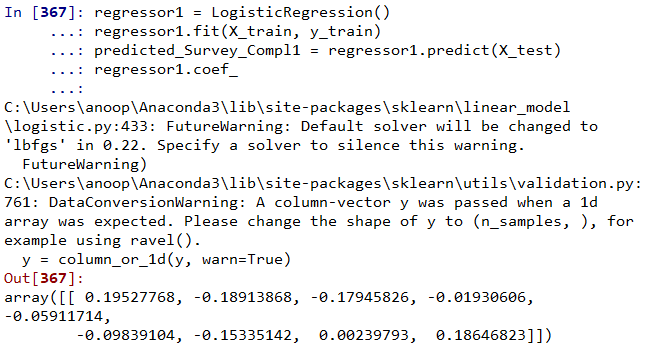
From the above confusion matrix, we can say that the accuracy rate of the model is  
 (418+406) / (418+302+312+406) = 824 / 1438 = 57. 3%

Hence the inaccuracy predicted by the model is = (1 - 57.3)= 42.7%

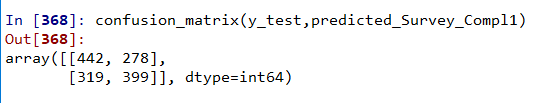
As such this is comparatively a good model.

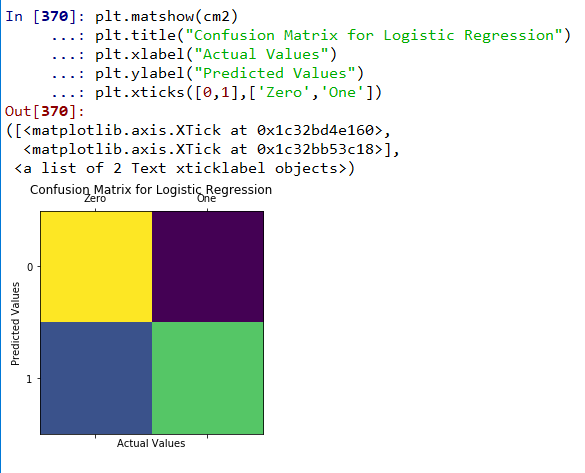
## Logistic Regression Model

As we are trying to look at the likelihood of an event logistic regression is by default a model to check as it uses the likelihood technique to optimize the model. Below are results of this regression model.



**Confusion Matrix:**





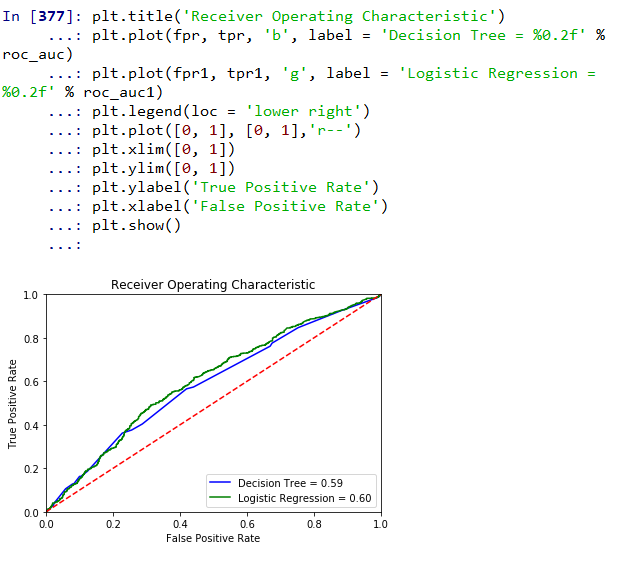
Results and Interpretation

The coefficients of the model suggests the below:

* Log Odds ratio of an Alumni completing the survey decreases by 0.179 when they belongs to Air Force. Which mean that Air Force veterans are less likely to fill the Alumni Survey
* Log Odds ratio of an Alumni completing the survey decreases by 0.189 when they belongs to Army. Which mean that Army veterans are less likely to fill the Alumni Survey.
* Log Odds ratio of an Alumni completing the survey decreases by 0.019 when they belongs to Coast Guard. Which mean that Coast Guard veterans are less likely to fill the Alumni Survey.
* Log Odds ratio of an Alumni completing the survey decreases by 0.059 when they belongs to Marines. Which mean that Marines veterans are less likely to fill the Alumni Survey.
* Log Odds ratio of an Alumni completing the survey decreases by 0.098 when they belongs to Navy. Which mean that Navy veterans are less likely to fill the Alumni Survey.
* Log Odds ratio of an Alumni completing the survey decreases by 0.015 when the veteran is Male.
* Log Odds ratio of an Alumni completing the survey increase by 0.002 when the veteran’s service period increases by a month.
* Log Odds ratio of an Alumni completing the survey increase by 0.186 when the veteran’s Pay Rank increases by 1.

# Model Assessments

**ROC Curve:**



From the ROC curve both Decision Tree and Logistic Regression models are almost equally good. they have values 0.59 and 0.60. Which indicates that the Logistic regression is slightly having edge over the Decision Tree model. The same is evident with the Green curve has more area covered under the curve.

References

<https://www.teradatauniversitynetwork.com/Community/Student-Competitions/2019/2019-Data-Challenge/Data-Challenge-Datasets>

<https://www.hireheroesusa.org/>

<https://www.federalpay.org/military/army/ranks>