**Analyzing the significance and impact of services provided by HHUSA on Hire Statistics of Job Seekers**

MSIS 5223: Programming for Analytics

**Anushree Joshi (20172800)**

**Manjusree Paimagham (20157161)**

**Devarshi Tharwala (20159739)**

**Ashutosh Chandra Sakinala (20167489)**

**Table of Contents**

[Executive Summary: 3](#_Toc7969859)

[Statement Of Scope: 3](#_Toc7969860)

[Project Schedule: 5](#_Toc7969861)

[Data Preparation: 6](#_Toc7969862)

[1) Data Access: 6](#_Toc7969863)

[2) Data Cleaning: 6](#_Toc7969864)

[3) Data Consolidation: 8](#_Toc7969865)

[4) Data Transformation: 8](#_Toc7969866)

[5) Data Reduction: 10](#_Toc7969867)

[6) Data Dictionary: 11](#_Toc7969868)

[Descriptive Statistics: 13](#_Toc7969869)

[Modeling Techniques: 26](#_Toc7969870)

[Assumptions: 26](#_Toc7969871)

[Model Goals: 27](#_Toc7969872)

[Data Splitting and Subsampling: 28](#_Toc7969873)

[Model Building: 30](#_Toc7969874)

[Logistic Regression: 30](#_Toc7969875)

[Decision Tree: 31](#_Toc7969876)

[Model Assessment: 33](#_Toc7969877)

[Final Model-Conclusion: 37](#_Toc7969878)

# Executive Summary:

There are around 20.4 million U.S.A veterans in 2016, according to data from the Department of Veterans Affairs, representing more than 10% of the total U.S.A adult population. With U.S.A being the third largest military in terms of active personnel, employing 1347300 citizens in its ranks, the number of veterans adding to the current population is only increasing by the year. Most of the veterans are adults between the ages of 23 and 40. This poses a major problem in terms of reemployment of the personnel. Hire Heroes U.S.A is an organization that provides services that aid the veterans in getting hired. Our analysis is to identify the services that have a significant impact on a veteran getting hired. We also build a model to recognize which services have the highest impact on the hire count. Through these insights the organization can optimize the budget allocation towards the services based on their contribution on improving the hire count.

# Statement Of Scope:

The goal of our study is to evaluate the effect of services provided by the Hire Heroes USA, which have a significant impact on hiring rates of the job seekers. This study includes records of 105744 job seekers derived from the Salesforce database which is the Customer Relationship Management (CRM) for Hire Heroes USA.

We will be evaluating the following:

1. The effects of participating in O2O certification program, VCTP certification program, newsletter, Job Board, resume review, updated resume review, HHUSA workshop, virtual workshop, career fair, virtual career fair, webinar and volunteer assistance.
2. Perform a comparative study to determine which service contributes the highest to the hiring count of HHUSA.

**Predictor and Target Variable:**

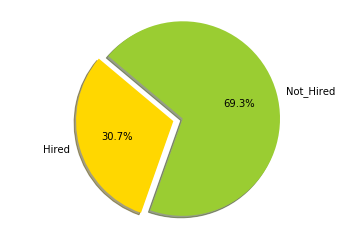
The target for our study is to predict HiredStatus of the job seeker which is a binary categorical variable consisting of 0’s and 1’s, where 1 indicates a hire. Our objective is to predict the hiring probability of a job seeker using:

* O2O certification program (O2O\_Certification)
* VCTP certification program (VCTP\_Certification)
* Newsletter (Newsletter)
* Job Board (JobBoard)
* Resume review (ResumeReview1)
* Updated resume review (ResumeReview2)
* HHUSA workshop (OfflineWorkshop)
* Virtual workshop (OnlineWorkshop)
* Career fair (OfflineCareerFair)
* Virtual career fair (OnlineCareerFair)
* Webinar (Webinar)
* Volunteer Assistance (VolunteerServices)

The target variable is coded as

     HiredStatus = 1; if the job seeker is hired

     HiredStatus = 0; if the job seeker is not hired



# Project Schedule:

We are planning to complete the project in approximately 97 days. We scheduled a weekly team meeting for a duration of 4 hours each, to discuss about the work progress and what should be done next. Work is done as a group task during the meeting sessions as well as divided among the team mates with each team mate responsible for their individual assignments. The individual assignments are reviewed by the team followed by integration of the code to the master script once verified. In case any team mate encounters issues whilst executing his assignment, we conduct online troubleshoot session to resolve the issue.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Research Task** |  | January | |  | February | | | |  | March | |  | April | | | | May | |
| W  1 | W  2 | W  3 | W  4 | W  1 | W  2 | W  3 | W  4 | W  1 | W  2 | W  3 | W  4 | W  1 | W  2 | W  3 | W  4 |
| 1. Project Proposal – Working Idea |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2. Statement Of Scope |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 3. Data Collection |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 4. Data Consolidation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 5. Data Cleaning |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 6. Data Transformation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 7. Data Dictionary |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 8. Executive Summary |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Finalization and Submission of Deliverable I** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1. Model Selection |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2. Build Model |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 3. Assess Model |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Compile Deliverable I and II and Submit** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1. Compile Final Report |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2. Presentation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 3. Record Presentation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Final Deliverable: Presentation** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

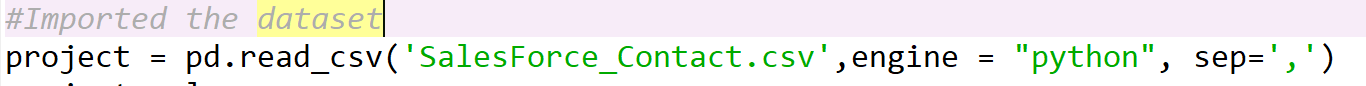
# Data Preparation:

## 1) Data Access:

We have obtained the data through the Teradata University network’s annual data challenge where they release a data set from which the participants have to analyze and submit their findings and results. The data set made available this year is quite interesting as it relates to the veteran recruiting services. The recruiting platform, Hire Heroes USA, uses salesforce as a Customer Relationship Management tool, hence most of their data is structured based on salesforce. The original data set has a total of 132446 observations (Rows) and 391 dimensions (Columns).

Data Source:

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



## 2) Data Cleaning:

1. **ClientOrOthers**

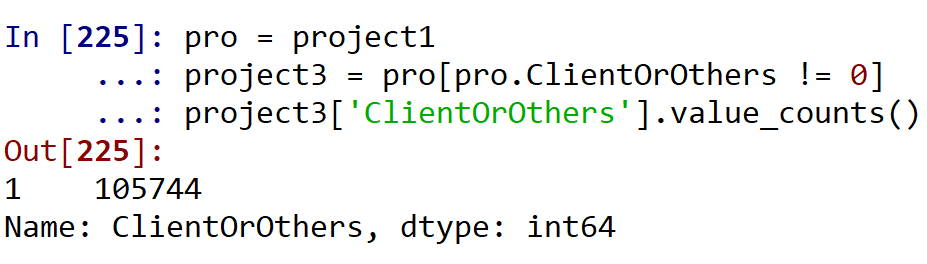
This is a categorical variable which indicates the number of registrants who are job seekers as 1 and non-jobseekers such as donors or volunteers as 0. We have

  Jobseekers/Clients or 1 = 105744

  Others or 0 = 26701

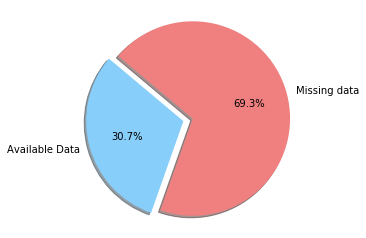
As our business question revolves around the hired status of the job seekers, we require only job seekers records and the corresponding usage of services. So, we have filtered our dataset with records for

  ClientOrOthers = 1



1. **OnlineTraining**

The online training indicates the people who underwent online training provided by Hire Heroes USA. After going through the variable description, we found a major portion of the data as missing that is 69.3% and 30.7% is available.

****

As the major amount of data is missing, we dropped the attribute so as to avoid biased results.

1. **Handling missing value**

Following are the number of null records in some columns:

JobBoard = 158

ResumeReview2 = 159

OnlineCareerFair = 171

OnlineWorkshop = 171

Webinar = 171

VolunteerServices = 6304

As the number of null records in Volunteer Services is 6304, which is very less compared to the total data set of 105744 records, we dropped the missing values from the data set. Also, we can see that the remaining columns has negligible amount of null values compared to the total data set, we did not refrain from deleting those null values as it will hardly affect the predicted result.

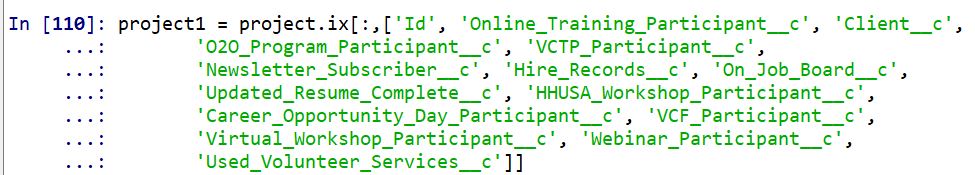
*project3.isnull().sum()*

*project3 = project3.dropna()*

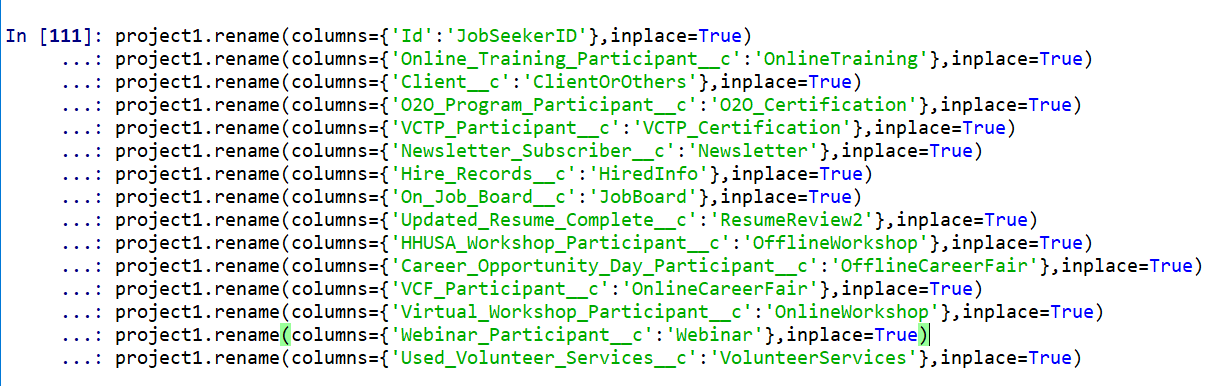
## 3) Data Consolidation:

Out of 13 datasets from the Teradata challenge, we have chosen only one dataset i.e Sales force contact. This dataset has 391 columns with 132445 records from which we have used 15 columns according to the need of our business question.

So, we have created a separate dataset for those 15 columns which has 132445 records.



We have renamed all those 15 columns and the code for that is below.



## 4) Data Transformation:

As our business question targets the hired status and wants to predict the hired status of the job seeker, we have formed a new attribute named ‘HiredStatus’.

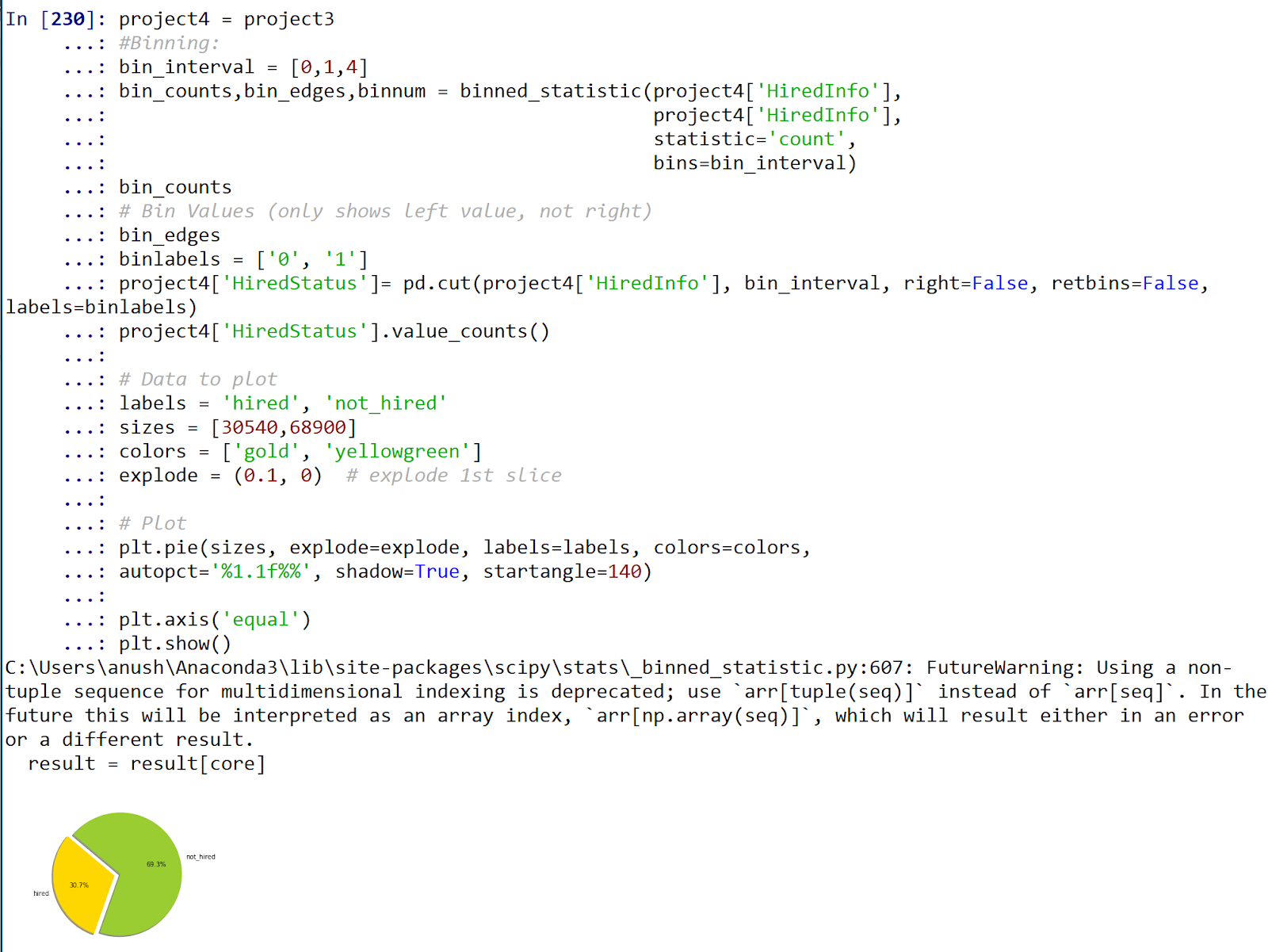
  HiredStatus = 0 if the job seeker doesn’t get hired

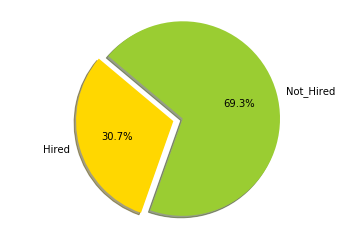
HiredStatus = 1 if the job seeker gets hired

So as to create this new attribute, we have binned the available attribute ‘HiredInfo’. HiredInfo indicates the number of hires the job seeker has got. So binned the attribute according to the following intervals:

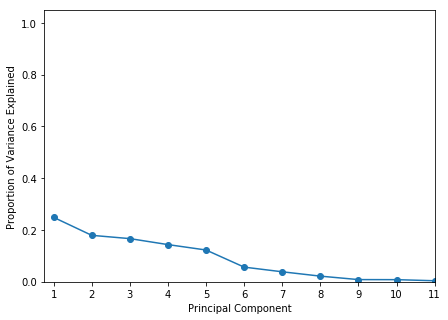
  HireInfo = 0 => HiredStatus = 0 or Not Hired

HireInfo = 1 or more  => HiredStatus =1 or Hired





## 5) Data Reduction:

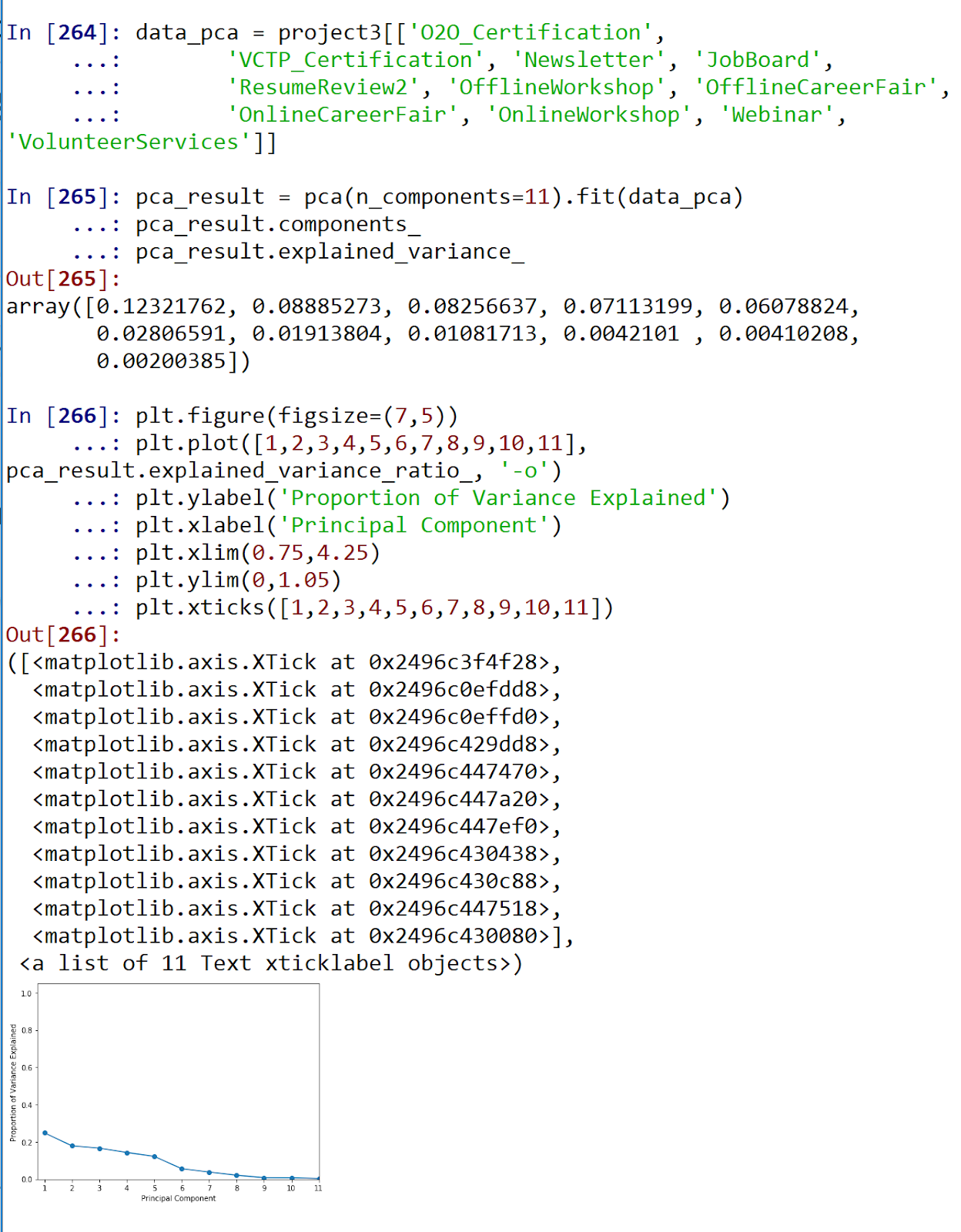


**Principal Component Analysis:**

PCA or Principal Component Analysis is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The technique that creates linear combination of columns. This process leads to creation of new columns.

In our PCA, we have found through the scree plot that we are not dealing with 11 but 8 significant dimensions. As we can see in the scree plot, that the graph starts flattening after 8 components out of 11 components.

Below we have the snippet of the code used to create a data frame consisting of the 11 variables, on which we perform PCA to identify the correlation. We follow this to generate a scree plot in order to get a better understanding of the correlation.



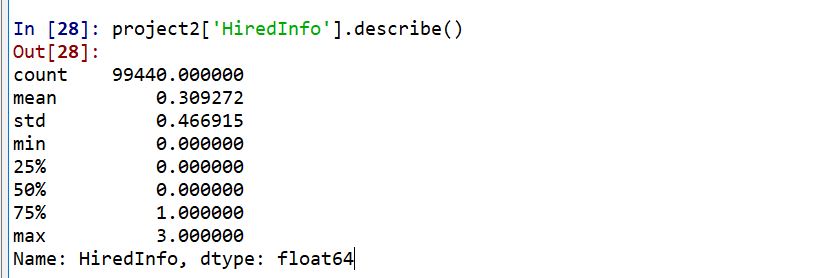
## 6) Data Dictionary:

Below table shows the attributes names of the data set adapted in the project, with their description and data type

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Type** | **Example** | **Description** |
| **JobSeekerID** | object | 0030z00002PrWd6AAF | ID for the job seekers |
| **ClientOrOthers** | int | 1 | Is the contact a Client (Job Seeker) or others ( Volunteer or Donor) |
| **O2O\_Certification** | int | 0 | Indicates job seeking client is also pursuing certification through the O2O program |
| **VCTP\_Certification** | int | 0 | Indicates client is pursuing professional certification through VCTP program |
| **Newsletter** | int | 1 | Client does / not receive the HHUSA newsletter |
| **HiredInfo** | int | 0 | Number of confirmed assisted hires with unique records |
| **JobBoard** | float | 0.0 | Indicates client has created a profile on HHUSA job board |
| **ResumeReview2** | float | 1.0 | Typically used for returning clients - further revisions to client resume beyond initial HHUSA resume review / development |
| **OfflineWorkshop** | int | 0 | Indicates job seeking client has attended a workshop hosted by HHUSA |
| **OfflineCareerFair** | int | 1 | Client has attended/participated in the career fair |
| **OnlineCareerFair** | float | 0.0 | Client has attended / participated in the virtual career fair |
| **OnlineWorkshop** | float | 1.0 | Client has attended one or more virtual workshop events |
| **Webinar** | float | 1.0 | Client has attended one or more webinar events |
| **VolunteerServices** | float | 0.0 | Participated in a mentoring session or mock phone interview |
| **HiredStatus** | Category | Hired/Not Hired | Indicates whether the person or hired or not |

# Descriptive Statistics:

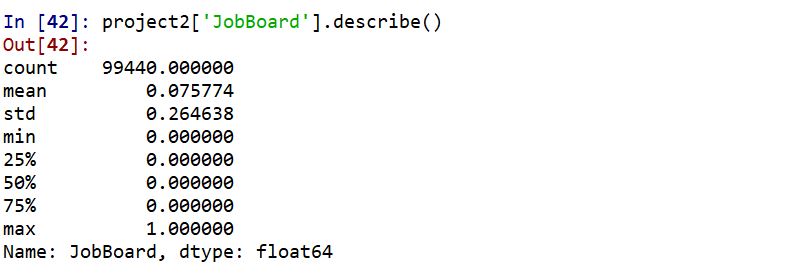
Before we start with our predictive analytics and run the model, let us study few of the variables:

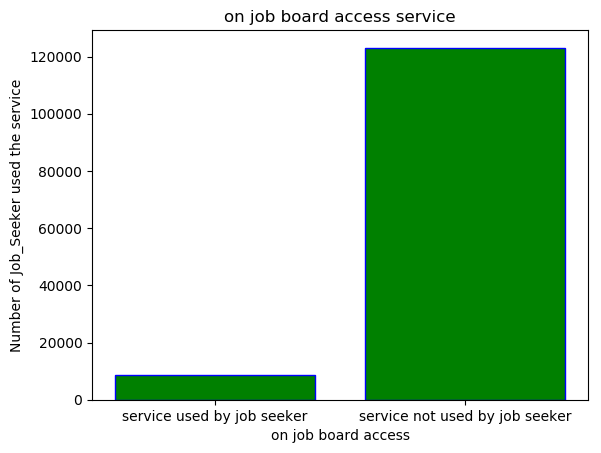
1. **HiredInfo**

Minimum value is 0 and the maximum value is 3. 0 indicates that the person is not hired and 1, 2, 3 indicates number of confirmed hires with unique records. So, we have binned 1,2,3 as 1 indicating the hired status of a job seeker.

1. **JobBoard**

This variable will tell us how many job seekers have used job board to register and look for the jobs.

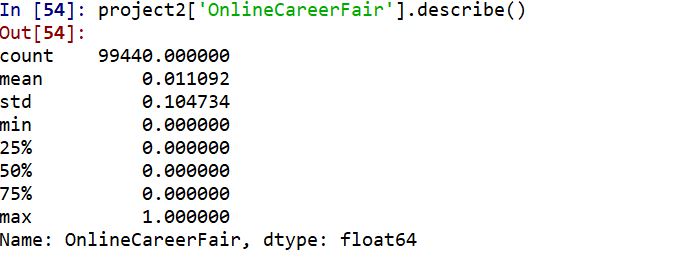


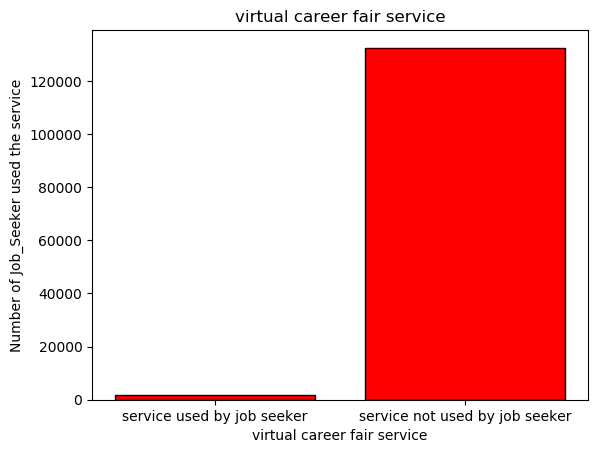


The mean of job board says that 75% of the people have registered in the job board.

1. **OnlineCareerFair**

This variable is used to estimate the number of people made used of this service.

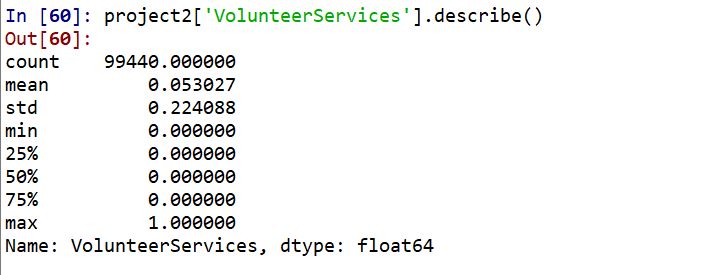


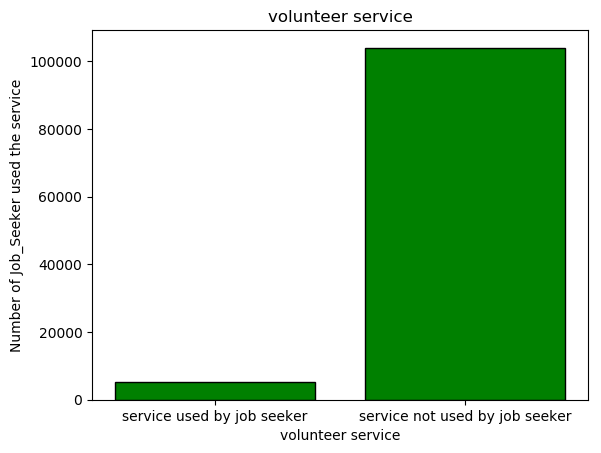


We can see from the graph that most of the job seekers did not make use of the online career fair. Out of 99,440 job seekers, only 1103 participated in the online career fair.

1. **VolunteerServices**

Volunteer Services will tell us how many people have used the mentoring session or phone mock interview.

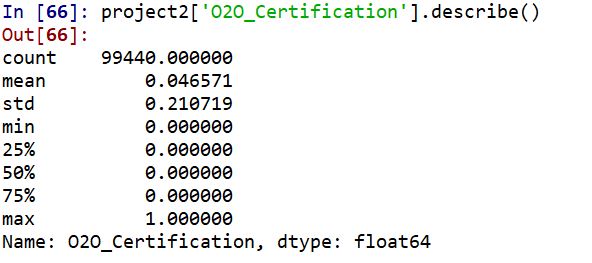
****

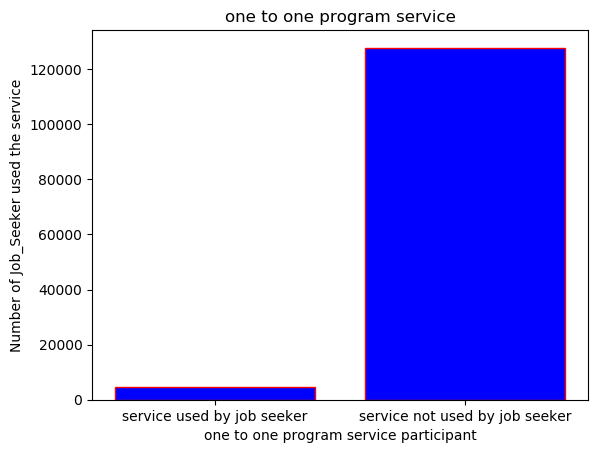


The average number of people used this volunteering service is very less. Only 5273 people have used this service out of 99,440.

1. **O2O\_Certification**

This variable gives information about the number of people participated in onward to opportunity(O2O) certification which helped them in getting hired.

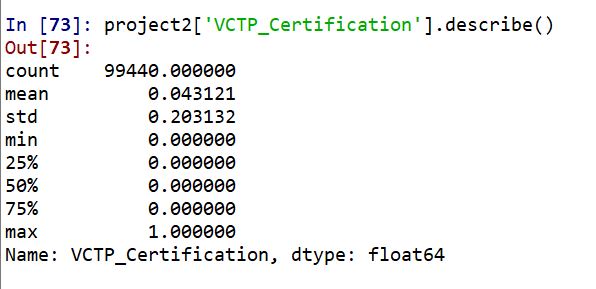


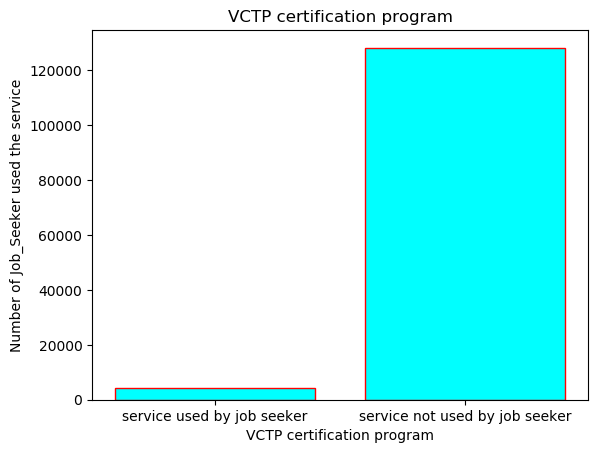


4631 job seekers did the O2O certification program and the remaining 94809 did not do it.

1. **VCTP\_Certification**

This Variable will indicate how many clients are pursuing the professional certification through VCTP program.

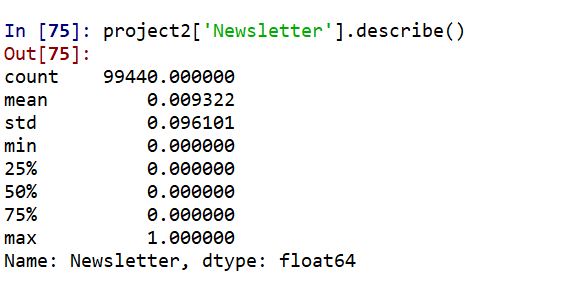


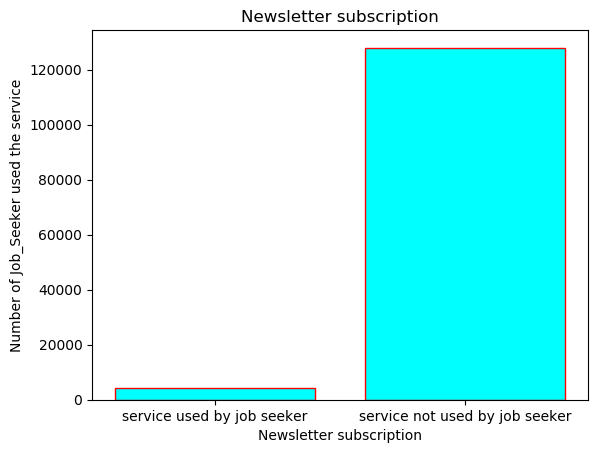


Only 4288 clients have participated in VCTP certification.

1. **Newsletter**

This variable will tell us how many clients receive the newsletters from Hire Heroes USA (HHUSA).

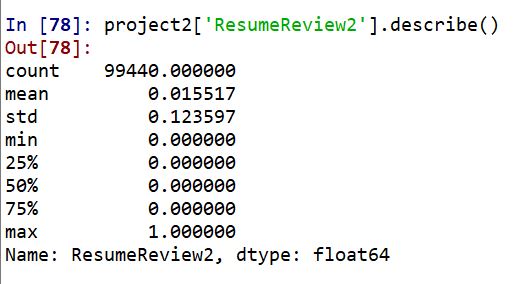


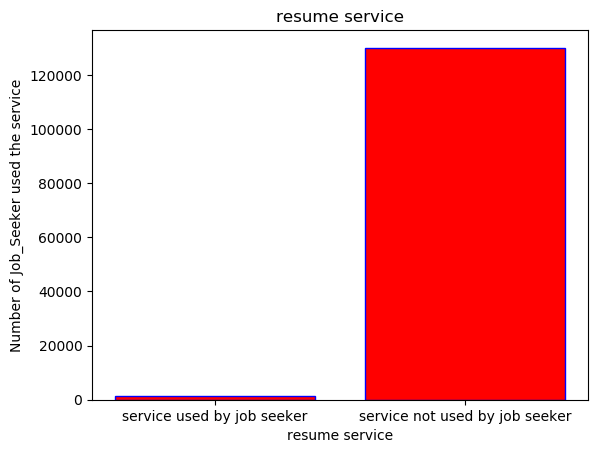


The average number of people receiving the newsletters from HHUSA are very few. Only 927 people receives newsletters from HHUSA.

1. **ResumeReview2**

This variable indicates how many clients have returned for the second time resume review beyond initial HHUSA resume review.

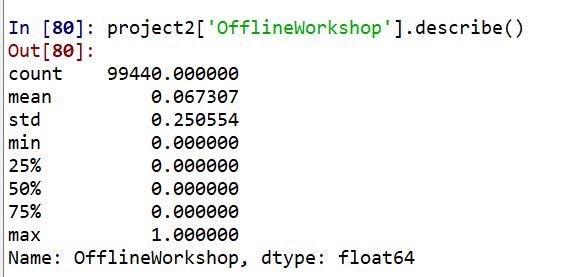


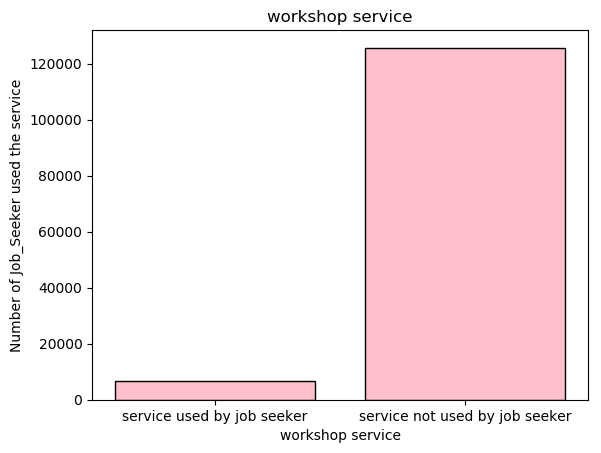


1543 out of 97897 people have returned for the second time resume review.

1. **OfflineWorkShop**

This will tell us how many people have attended the workshop hosted by HHUSA.

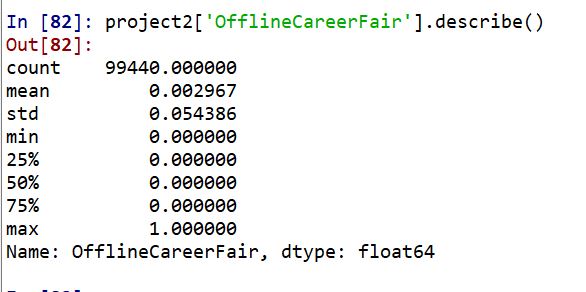


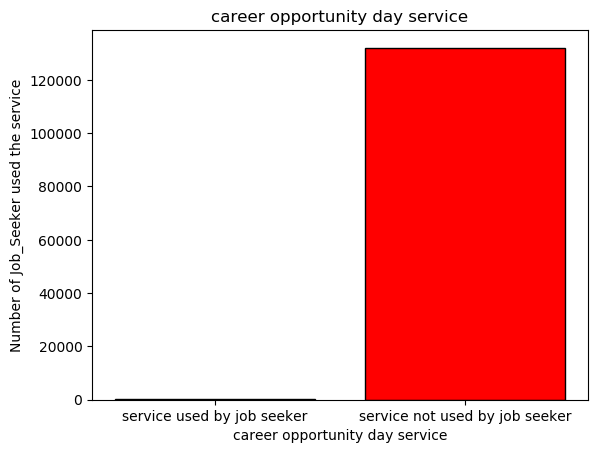


6693 job seekers have participated in the workshop hosted by HHUSA.

1. **OfflineCareerFair**

Offline Career Fair will tell us how many job seekers have attened the career fair offline.

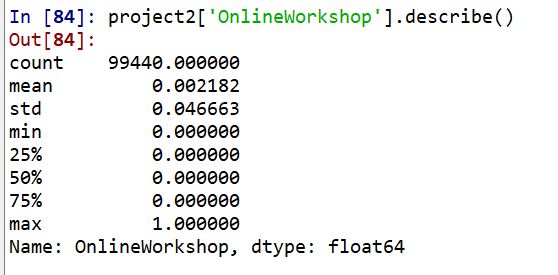


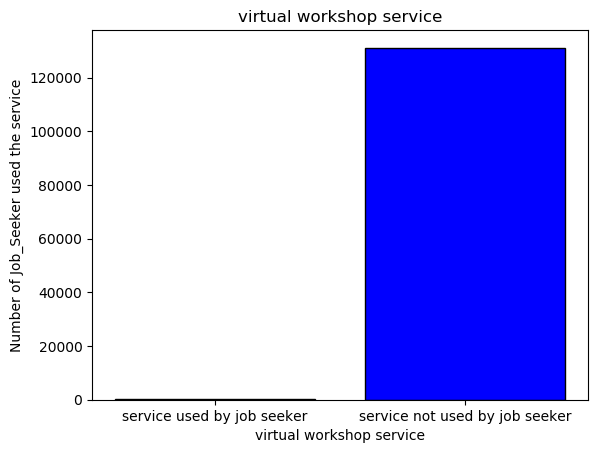


Only 295 people have made use of this service.

1. **OnlineWorkShop**

This will indicate how many clients have participated in one or more online workshops.

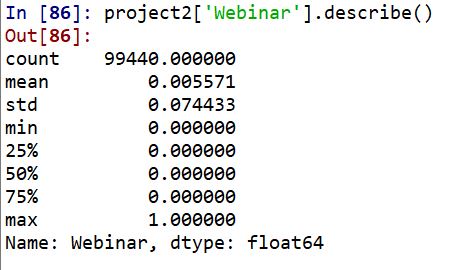


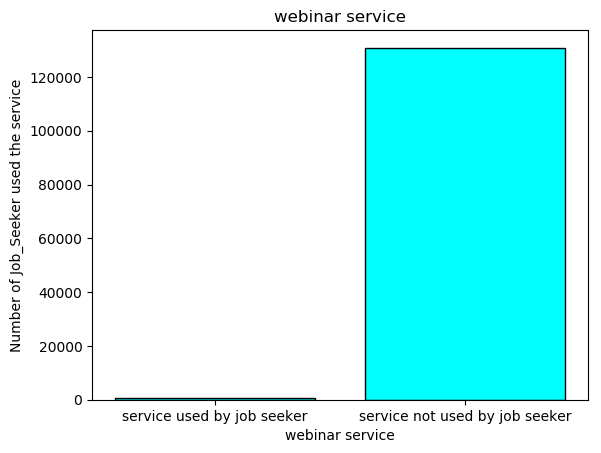


A total of 217 have participated in one or more online workshops.

1. **Webinar**

It is used to identify the number of clients who have attended the webinar sessions.

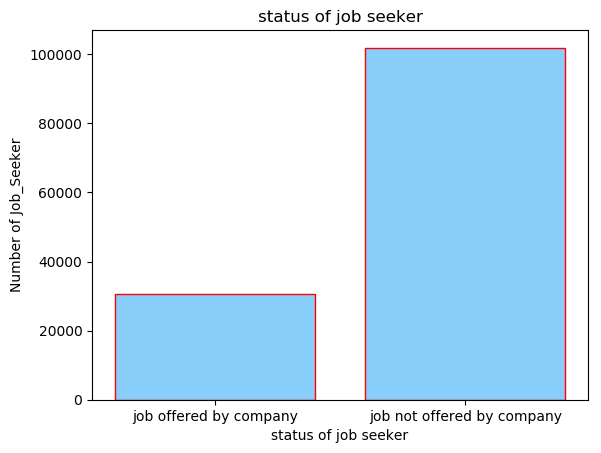




554 clients have made use of this webinar session.

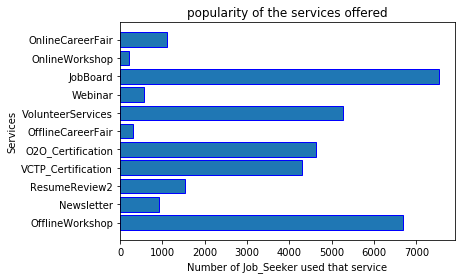
1. **HiredStatus**

This indicates the number of people who got hired.



By using all the above-mentioned variables, we can find out the impact of each variable on the predictor variable which is HiredStatus. We are going to analyze how much each variable is going to contribute a job seeker in getting hired.

-The services as per their popularity of use are as below;



# Modeling Techniques:

After the data has got cleaned and processed, we can start with the modelling techniques to predict the correlation between the hiring rate and the services by the Hire Heroes USA. Also, we have to see which variable has the highest significance in the hiring rate. Our target variable is Hiring Status which is a binary variable with values 0 for not hired and 1 for hired. Therefore, we would go for classification techniques. We are choosing decision tree and logistic regression.

We will be using logistic regression to model our data since it examines the relationship between multiple independent variables and the log odds by calculating changes in the log odds of the dependent as opposed to the dependent variable itself. The assumption here is log odds of Hiring status is linearly related to independent variables.

Next modelling technique we aim to build is Decision tree, since it gives a visual retreat of decision situation. The branches of the tree would clearly draw a distinction and would give a clear picture about which independent variables determine the probability if a job seeker gets hired and the order of their importance as well.

# Assumptions:

**LOGISTIC REGRESSION:**

A logistic regression model is used to predict the outcome when the response variable is categorical. In our model, the target variable is ‘Hired Status’ which is a binary categorical variable consisting of 0’s and 1’s. We assume the model would work with high accuracy because of the ample amount of observations (100,000). We also have very few Outliers which goes will with impacting the skewness.

The first assumption here is that our independent variables have some relation with the target variable. These variables might have varying effects on the response variable. We have more than one independent variable in our model and hence it would be a multiple logistic regression model. This would help us in finding the combined effect of all the independent variables such as the 'O2O\_Certification', 'VCTP\_Certification', 'Newsletter', 'JobBoard', 'ResumeReview1' and 'ResumeReview2' on HiredStatus.

Logistic Regression assumes linear relationship between the logit of independent variables and dependent variables, but however there need not be the assumption of linear relationship between the actual dependent and independent variables. Normality is not a concern with respect to the target variable and Homoscedasticity isn’t an issue with respect to the levels of independence.

**DECISION TREE:**

The main assumption to consider decision tree model is that there is no necessity of having a relationship between dependent variable and independent variables. We also assume the model would work with high accuracy because of the ample amount of observations (100,000).

We also have very few Outliers which goes will with impacting the skewness. Also, decision tree provides more appropriate results for the dataset that has large amount of observations. As our project has a huge data set with binary target variable HiredStatus and independent variables of categorical type and only in the format of binary as in 0’s and 1’s, so there is no problem in choosing decision tree to be executed.

# Model Goals:

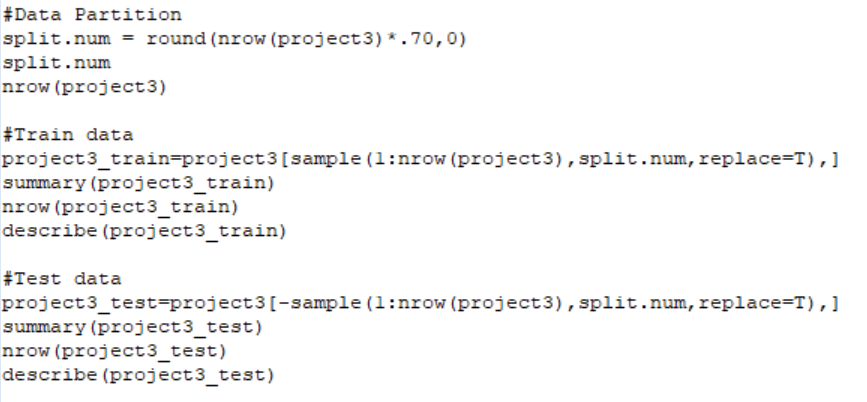
**LOGISTICAL REGRESSION:**

The goal of the logistic regression is to find the hiring status of job seeker which is represented by the HiredStatus and how it will be effected with every unit raise in the independent variables 'O2O\_Certification', 'VCTP\_Certification', 'Newsletter', 'JobBoard', 'ResumeReview1' and 'ResumeReview2'. We have considered these categorial variable with binary value in order to predict the effect and significance of the independent variable on predicting the HiredStatus of the job seeker.

**DECISION TREE:**

The goal of the decision tree is to build a model that predicts the value of HiredStatus by creating simple decision rules drawn from the categorical data with binary values. The data features especially from the independent variables. Decision tree comes up with pretty simple rules about how the predictor variables affect HiredStatus. We consider the splitting criteria based on the significance of the independent variable from the highest to lowest alongside pruning to get the most relevant results. The decision tree is simple to understand even for the people who are not aware of the subject, alongside giving a case by case representation of the effect of the services provided by Hire Heroes USA after removing the variables that cause the inter variable correlation, on the HiredStatus.

# Data Splitting and Subsampling:



Data splitting is generally the act of partitioning the available data into two portions, usually for cross-validation purposes. Cross-validation techniques belong to the conventional approaches where we ensure good generalization and to avoid overtraining. The basic idea is to divide the dataset into two subsets, one is training, and the other is testing. Cross-validation techniques can also be used when evaluating and mutually comparing more models, various training algorithms, or when seeking for optimal model parameters Training portion of the data is used to build a predictive model as the model sees this set of data while determining the best data transformation and to determine which predictors to include in the model and which one to eliminate.

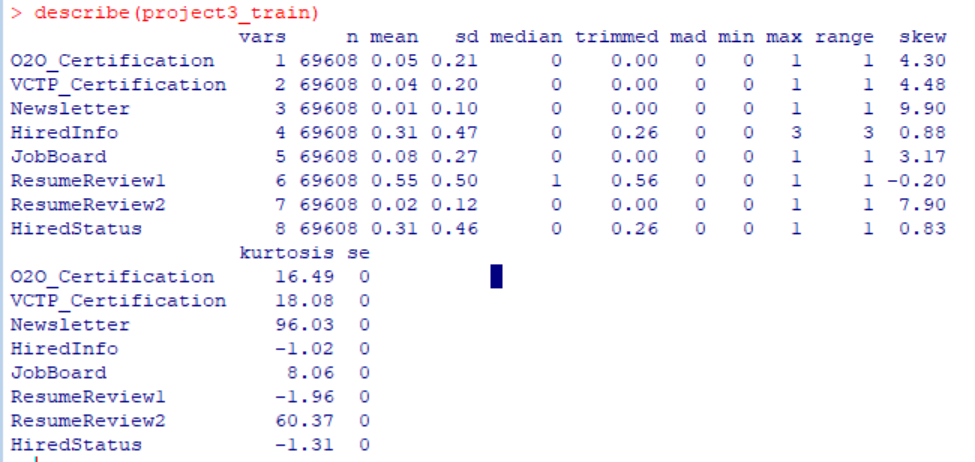
The training set is used only after the model is being built and it is used to compare predictive capabilities across different models and confident estimate evaluate the model’s performance. Here in our project, we are performing the data splitting and the proportion chosen is 70% for the training set and 30% for the test. The idea is that more training data is a good thing because it makes the classification model better while more test data makes the error estimate more accurate. Reason for choosing 70-30 instead of 50-50 is:

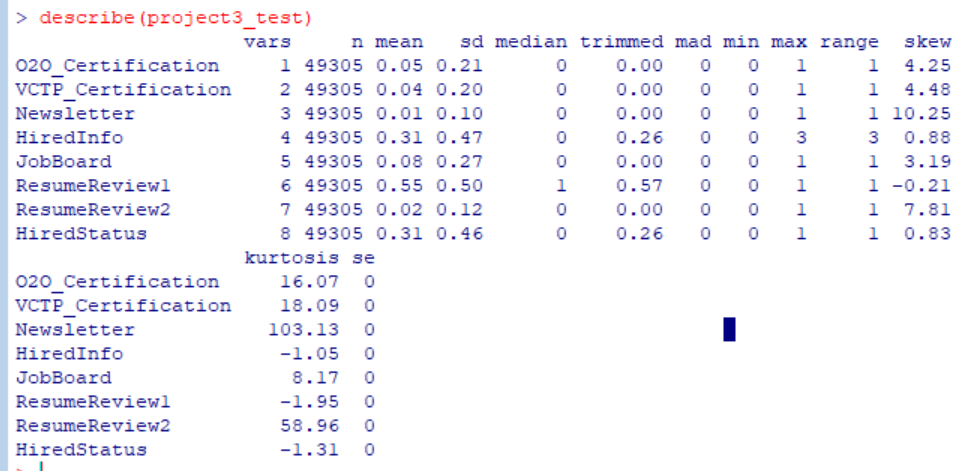
• When we take data set into consideration, we have huge dataset, so we prefer to choose 70- 30 split instead of 50-50 split.

• To determine the predictive ability, we choose the 70-30 division for training and testing dataset instated of 50-50.

• higher percent for training data because we want to assure that we have enough data so that we can identify properly the trends for our models as lower percent for training may not recognize the larger events for classification.

**Comparing the split datasets:**

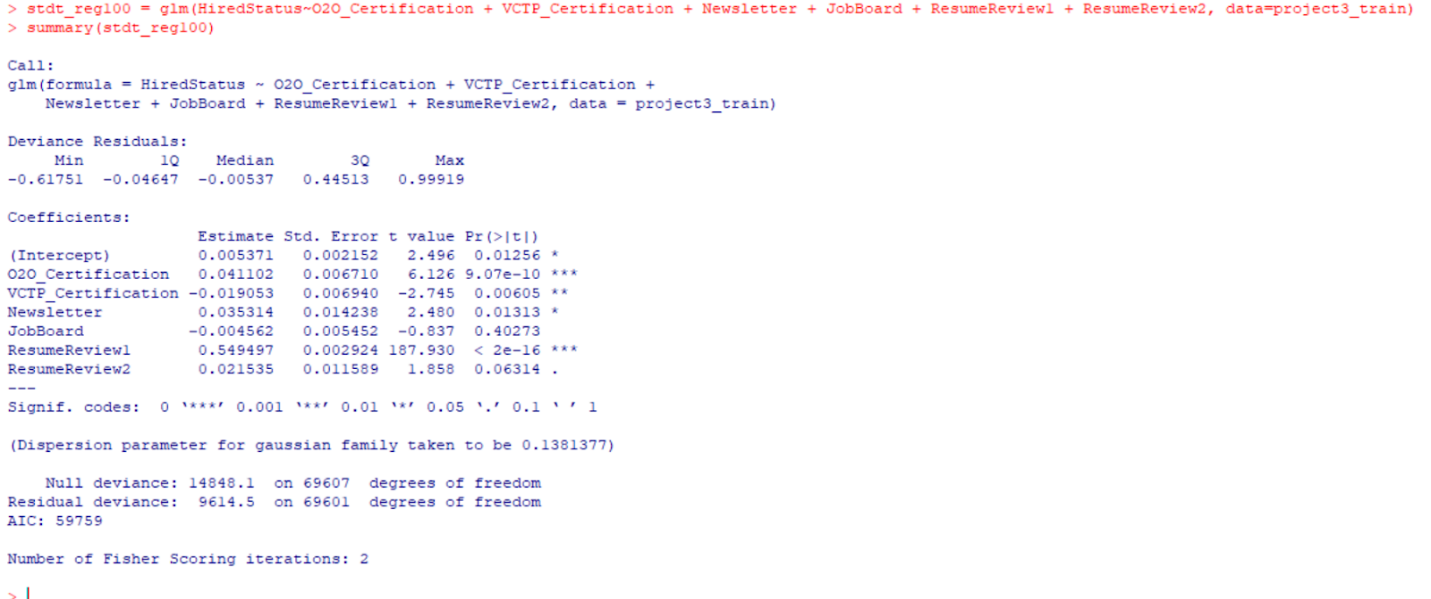




# Model Building:

## **LOGISTIC REGRESSION:**

Our dataset consists of a large number of independent categorical variables with binary class structures. Through this model, we get the significance of each predictor variable (Hire Heroes USA) Service in predicting the HireStatus. We also get the combined significance of the services to judge which services perform well together, or if a services effectiveness increases when paired with another.



**Results & Interpretation**

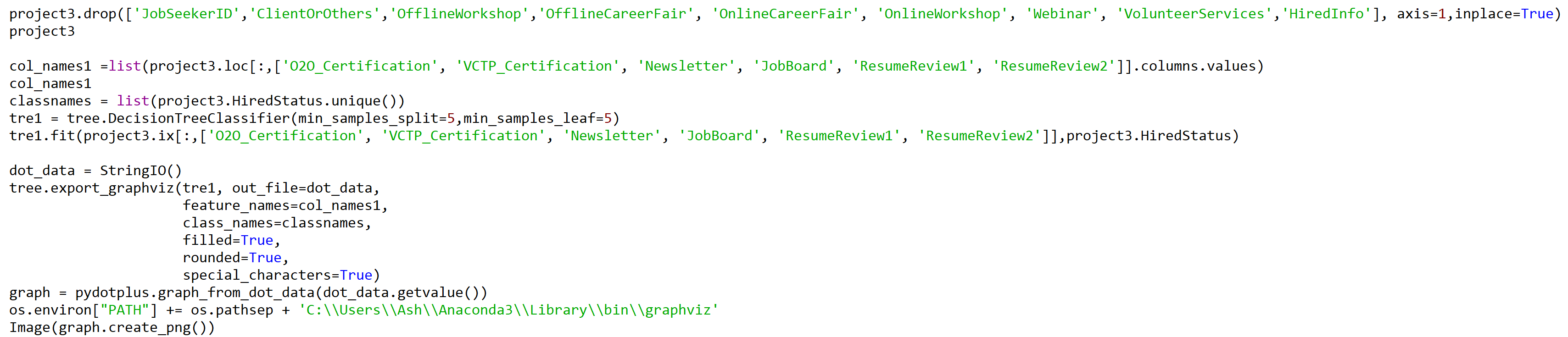
According to the logistic regression summary,

Log (odds ratio of jobseekers’ hired status as unhired) = 0.005371 +0.0041102\*(O2O\_Certification)-0.019053\*(VCTP\_Certification) +0.035314\*(Newsletter)-0.004562 \*(JobBoard)+0.549497\* (ResumeReview1) +0.021535\*(ResumeReview2)

The regression equation with positive coefficients indicates that there is a rise in the odds of the jobseekers’ hired status as unhired with subsequent increment in the explanatory variables listed in the equation. Similarly, regression with negative coefficients indicates that there is a fall in the jobseekers’ hired status as unhired with a subsequent decrease in the corresponding independent variables.

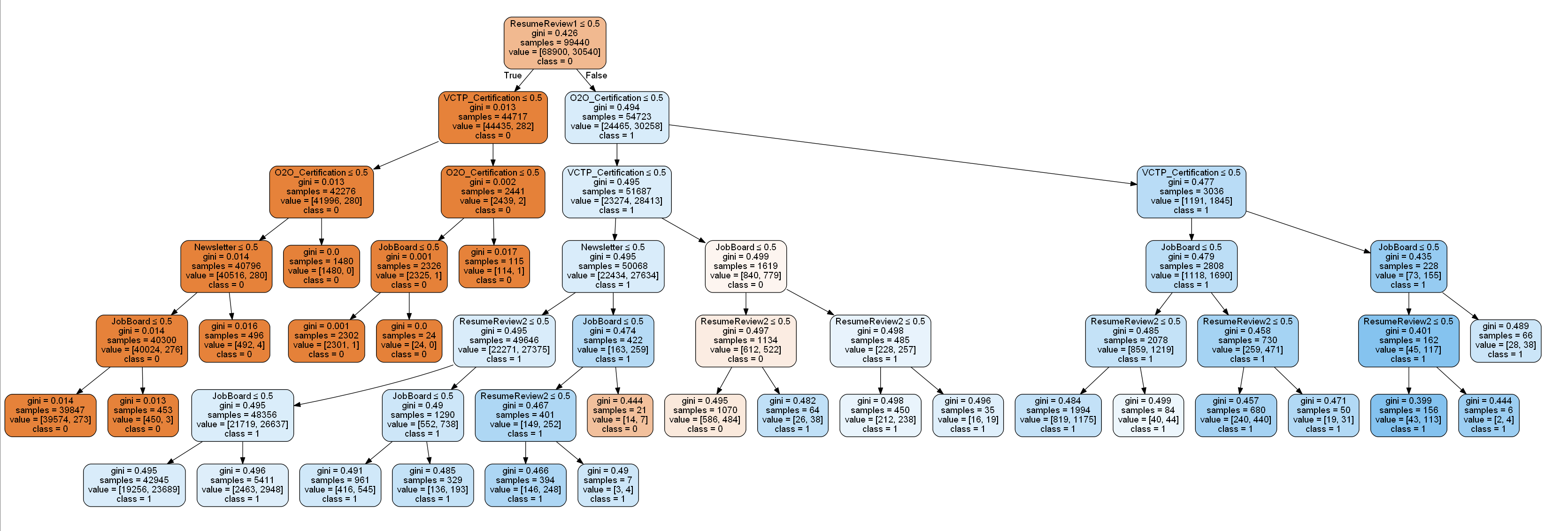
For example, from the above equation if O2O\_Certification increases by 1 unit, the log of odds that a jobseeker’s status is hired will rise by 0.041102 that is odds will rise by e^ (0.041102) units. Similarly, if VCTP\_Certification raise by 1 unit then a log of odds for the subscription of term deposit by customer falls by 0.019 that is odds will fall by e^ (0.019) units. Null deviance for the above model is 14848.1 on 69607 degrees of freedom, this is the deviance value that shows how dependent variable is predicted given the model only includes the intercept value and doesn’t include any explanatory variable. Similarly, residual deviance for the above model is 9614.6 on 69601 degrees of freedom, this is the deviance value that includes both intercept and explanatory variables in the model.

## **DECISION TREE:**



**Results & Interpretation**

Through the decision tree, we can see that the most significant predictor of job seekers recruitment into a company is Resume Review 1. Also, we see Resume Review 2 alone doesn’t add any value but when combined with Resume Review 1 it increases the probability drastically and this makes sense as Review 2 is to be conducted in sequence with Review 1.



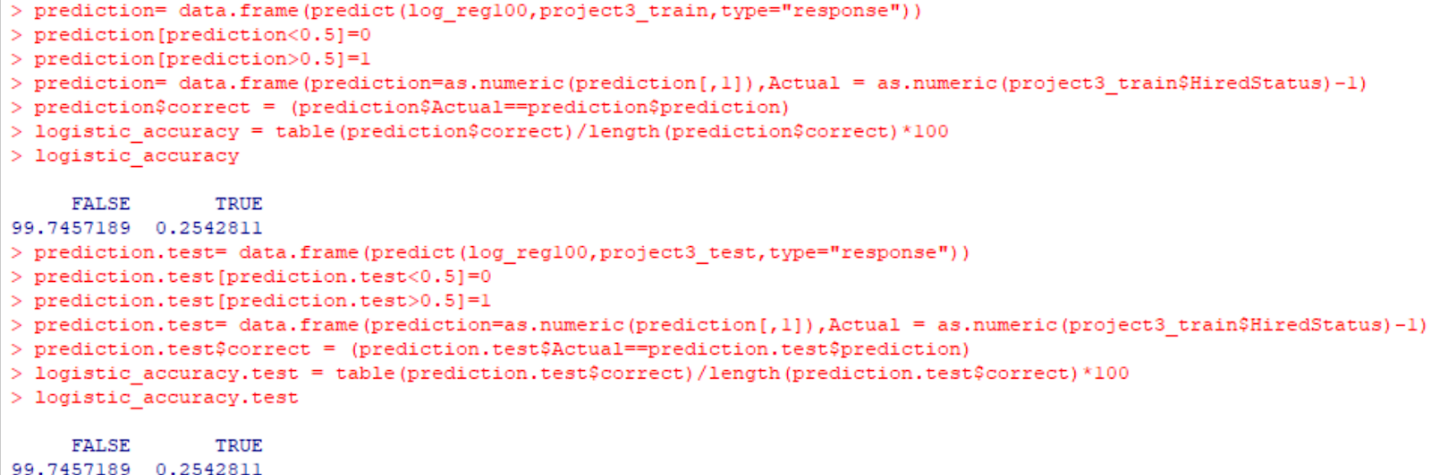
# Model Assessment:

**LOGISTIC REGRESSION:**

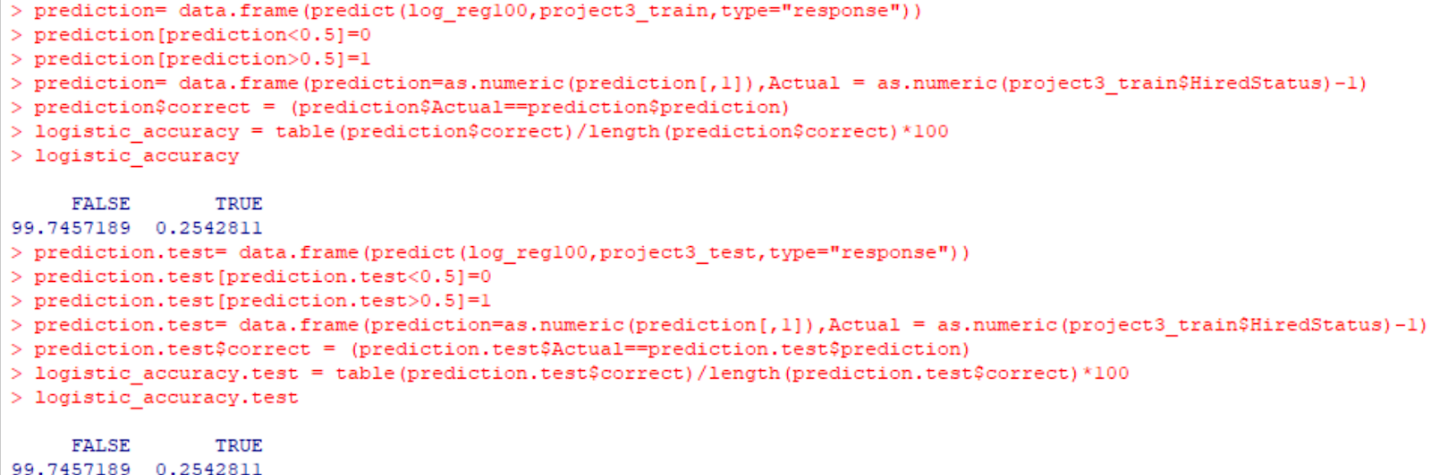
**Accuracy Measure:**

Accuracy of the model is a key metric which determines how well is the classification done by the model. Higher the accuracy means the model has higher predictive ability. The below code shows the code for determining the accuracy of the model on both training dataset and testing dataset.

Training:

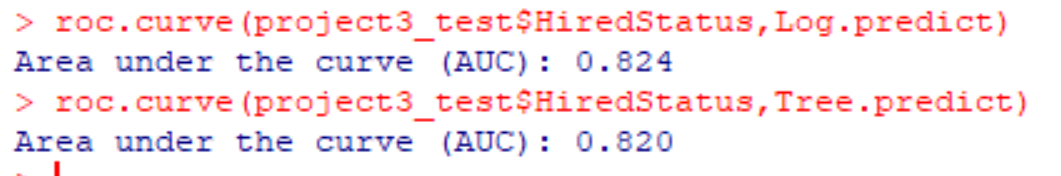
****

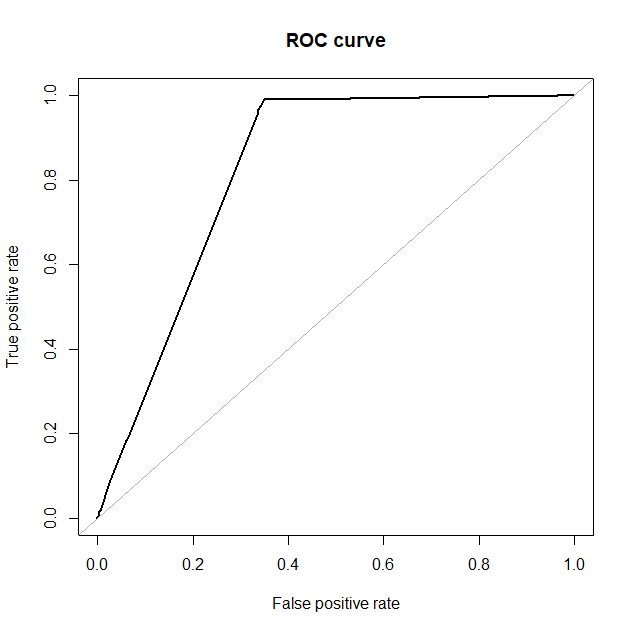
Testing:

****

**ROC:**

The ROC chart provides a means of comparison among different classification models. It is a plot between false positive rate i.e. (1-specificity) on the x-axis and true positive rate i.e. sensitivity on the y-axis. In the ideal case, if the curve climbs quickly towards top-left then it means that the model is performing well. The code below plots the ROC curve and determines the area under the curve.

****

****

**Strength:**

* Logistic regression doesn’t require any assumption of direct relationship between target variable and predictors.
* With the use of stochastic gradient, logistics model can be updated easily. Thus, many sceneries can be studied easily.
* The foremost important advantage of logistic regression is to calculate the odds of the dependent variable based on the weights of independent variable.
* Logistic regression can utilize feature of variable reduction/selection while creating the regression model by defining the entry/exit level while executing the model.
* Logistic regression model has high predictive power.

**Weakness:**

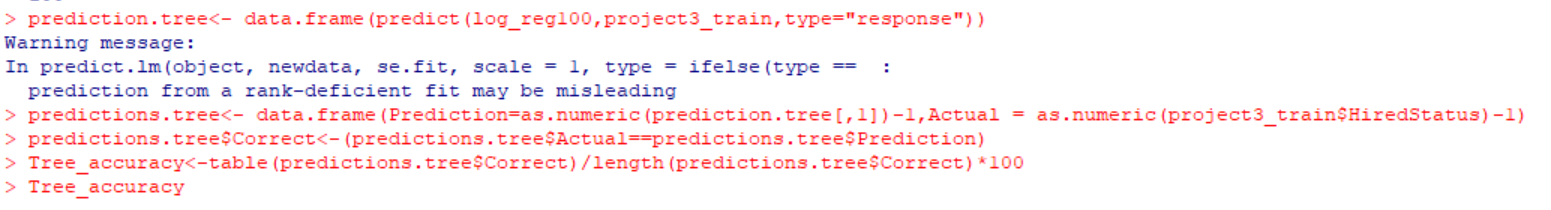
* Logistic regression is not appropriate for small sample sizes as it produces inaccurate parameter estimates.
* The model does not perform good when there is complex relationship between variables.

**DECISION TREE:**

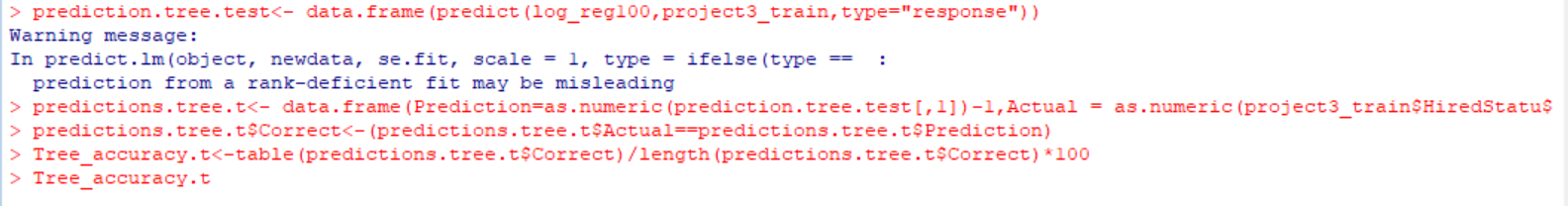
**Accuracy Measure:**

Accuracy of the model is a key metric which determines how well is the classification done by the model. Higher the accuracy means the model has higher predictive ability. The below code shows the code for determining the accuracy of the model on both training dataset and testing dataset.

Training:

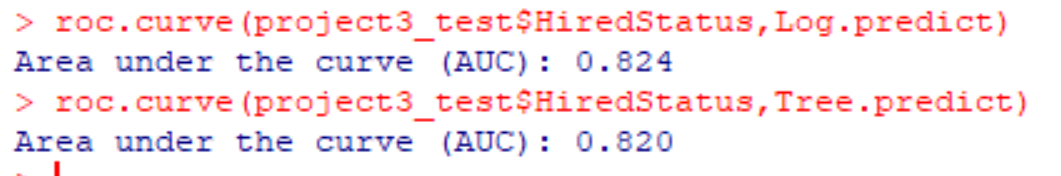
****

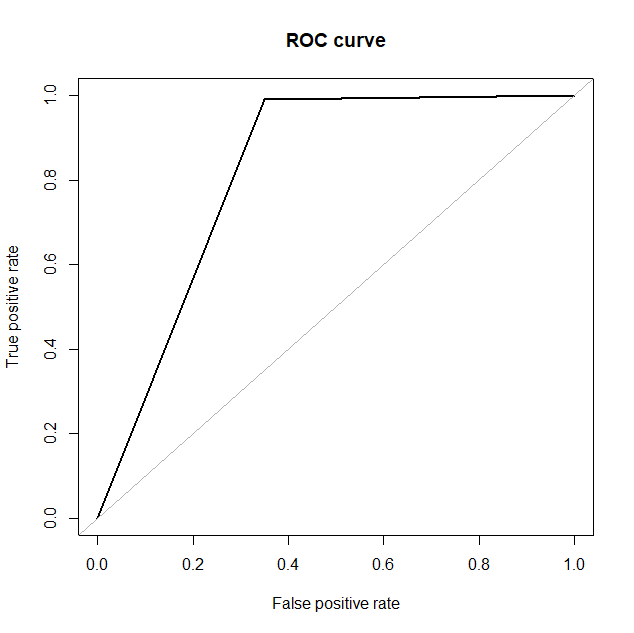
Testing:

****

**ROC:**

The ROC chart provides a means of comparison among different classification models. It is a plot between false positive rate i.e. (1-specificity) on the x-axis and true positive rate i.e. sensitivity on the y-axis. In the ideal case, if the curve climbs quickly towards top-left then it means that the model is performing well. The code below plots the ROC curve and determines the area under the curve.

****

****

**Strength:**

* Decision tree surely performs feature selection or variable screening. When we choose a decision tree to our dataset, variable having the most elevated summation of improvement value is fundamentally the most basic factor inside the dataset and feature selection is completed consequently.
* A decision tree requires modestly little effort from users for data preparation, to overcome the scale differentiates between.
* The result of the decision tree is not affected by the nonlinear relationship amongst variables. Thus, the scenarios in which there is non-linear relationship among the variable can be used and analyzed.
* One of the best feature of utilizing trees for analytics – simple to understand disclosed to administrators.
* The decision tree has a greater amount of accuracy and provides a case by case analysis of each combination of service.

**Weakness:**

* The short come of the decision tree model is that the more confusing the model turned into, the less valuable it is these leads to difficulty in measuring the importance of a variable node.
* As in our project we are dealing with the binary target variable, it doesn’t create any consequences but if it is continuous variable it is difficult to predict appropriately using decision tree especially when their range is high.

# Final Model-Conclusion:

We initially started off with the descriptive statistics that helped us in finding the summary statistics such mean, median, the standard deviation for numerical variables, number of levels, mode and etc. for categorical variables. Then we started with Logistic Regression. Logistic Regression provided the satisfactory results in which ‘ResumeReview’ had the highest significance according to the logistic regression model. such as McFadden’s R^2 has the value of 1.118, Accuracy of 89% on testing dataset and area under the curve (AUC) for ROC is 0.824.

In the next step, we proceed with the decision tree that helps us to identify the most important independent variables and thereby classify the ‘HiredStatus’. Results of the decision tree were quite satisfactory (Accuracy is 87% and AUC has value 0.820). Decision Tree could classify the HiredStatus into either Yes/No appropriately and we get to see a clear view of the significance and split levels of all the dependent variables.

So, from the above results, we can say that Logistic Regression can be obtained as a final model objectively as it has higher AUC Value. But the AUC values for both the models do not have a large margin of difference and subjectively, decision tree might be a better option because it gives us case by case prediction of the HiredStatus for the jobseekers with reference to the services. It also provides us with the services in the order of their significance alongside visualizing every possible combination of the services. We also pruned the tree to give the most relevant and most impactful nodes based on the degree of probability rise.