

**Empower U.S. military members, veterans, and military
spouses to succeed in the civilian workforce:
Using an Analytical approach**

**Course: DSBA/ MBAD 6211- Advanced Business Analytics
Instructor: Dr. Mousavi**

Table of Contents

Project Report	3
Introduction	3
Background	4
Data	5
Results & Discussions	21
Conclusions	21
Appendix	21
Reference	

1 Project Report

1.1 Introduction

Although most U.S. veterans transition to civilian life successfully, securing employment and reintegrating into civilian communities, some veterans face transition challenges that can lead to or exacerbate mental and physical health problems. Emerging research from a survey conducted by Prudential indicates that difficulty transitioning to civilian life is largely attributable to employment.

Hire Heroes USA as of February 1st, 2019 announced reaching a milestone of 30,000 veterans hired into jobs since its inception. The national nonprofit is a leader in the veteran employment space, providing personalized job search assistance to transitioning US military members, veterans and military spouses(Hire Heroes USA, Feb 2019). But still as reported in January 2019, the veteran unemployment rate was 3.7% and Women Veterans Unemployment: 2.7%. With an increasing veteran and military spouse population, it is critical to find the right job and be employed. The non-profit organization Hire Heroes USA, mission is to empowers U.S. military members, veterans and military spouses to succeed in the civilian workforce and their vision is become the Nation's preferred veteran employment service organization through a relentless focus on personalized career coaching that improves clients' quality of life and strengthens the U.S. economy.

Veteran poverty also becomes a concern due to continuous unemployment and difficulty to fit in the civilian society. The veteran poverty rate for vets between 18 and 34 years old is higher than all other age groups. This group, of course, means that the veterans of the Gulf War and Afghanistan have higher poverty rates than other veterans. Veterans are also using food stamps in larger numbers than ever before. Although the rate is still lower than for non-veterans, it is rising at a much faster pace. The Supplemental Nutrition Assistance Program (SNAP) has seen an unprecedented rise in usage over the last six years, and veterans have been applying to the program to make ends meet. Almost a million households with veterans living in them receive SNAP.

Only 17 percent of employers say veterans are viewed as strategic assets in the workplace, according to the survey, released by the marketing firm Edelman(Natalie Gross, July 2018) . And despite the large majority of veteran respondents saying they have education beyond a high school diploma, 46 percent of employers believe veterans do not pursue a college degree or vocational training. Applying for civilian jobs may be challenging for veterans as they attempt to translate military qualifications, skills, and experiences to civilian jobs (Hall, Harrell, Bicksler, Stewart, & Fisher, 2014; Harrell & Berglass, 2014). Based on a study conducted by The Center for a New American Security where they interviewed 87 individuals from 69 different companies to determine why companies refrain from hiring veterans, to get an employer perspective they determined the gaps such as skill translation, Skill mismatch, Negative stereotypes and Acclimation(Lisa Nagorny and Dan Pick, 2019).

Out of the several challenging questions that HHUSA raised, we felt that identifying the most important factors that impacts a veteran or spouse to get hired will help the organization the most since that is the primary objective of HHUSA. This analysis will help Hire Heroes analyse the importance of these gaps methodically and implement changes when supporting veterans. We are implementing data science techniques to determine the relationship between a client's demographic profile, amount of time spent working with individual clients (time to complete an assessment, time to complete resume, # of logged activities, etc.) and the time required for the veterans to get hired.

1.2 Background

A successful military veteran hiring program requires a strong foundation that runs much deeper than just getting the veterans recruited. It needs a phased-approach that begins with organizational planning and preparation to deploy the right applications, processes, and tools required to execute effectively and maximize business impact. However, most companies today do not have a clear picture of the number of veterans they already have employed, and lack of data on their current military workforce performance and the individual needs that may cause roadblocks for their military talent to reach full potential (by Jesse Canella, CEO – Military Talent Group and David Pollard, Chairman - PredictiveHR). Building the right strategy must first start with a comprehensive analysis and understanding of an organization's current workforce and the overarching business objectives.

Usage of data analytics in job placement is an area that is getting a lot of focus recently though this is fairly new compared to the use of predictive modeling in other areas such as retail and financial industries. Many of the firms and colleges are working on coming up with different models that help their clients and students to have a successful career opportunity. The problem that we are trying to address has two dimensions. While we want to know what are the factors that are influential in making a client succeed in their job hunt, this problem also wants to understand how applicable these variables will be for the military professionals' veterans and their spouses. There are a good number of studies that had happened on the former but predictive modeling or researches on the latter dimension seem to be very sparse. Similar predictive modeling that focuses on the successful career placement was conducted by Modern Education and Computer Science org (Ajay Kumar Pal, Saurabh Pal, July 2013) . A classification model was developed to predict the likelihood of a candidate being successful in job placement. As per the analysis, the team tried out different options such as Naïve Bayesian Classification, Neural Networks (Multilayer Perceptron) and Decision Tree with k-fold cross-validation. Another study conducted at the University of Nebraska([reference](#)) used the Probit Regression model to predict the outcome of a successful placement. While the likelihood is predicted using different classification methods, we are interested in determining which qualities or activities get the veteran hired through HH U USA portal. A classification model would be ideal for the prediction of the likelihood of a veteran to be hired over other veterans in the database. This analysis will be utilised to determine which military and post-military training and experiences were those veterans exposed to that assisted them in getting hired, can we replicate that experience by training the unemployed veterans.

Predictive analytics technology can play a vital role in this base lining exercise providing, for the first time in many cases, visibility into the value and opportunities that a focus on military talent reveals. This data-driven approach is unique way for an Hire Heroes USA organization to accurately measure the growth of veterans and thus impacting on community as a whole.

1.3 Data

The entire data for the Teradata Data Challenge is provided by a non-profit partner, Hire Heroes USA, whose mission is to empower U.S. military members, veterans and military spouses to succeed in the civilian workforce. Hire Heroes USA uses Salesforce as their CRM, so most of their data is structured based on Salesforce's use of Objects. The data provided includes 13 spreadsheets, which we will be using to answer our questions.

The files used are:

1. Contacts

This dataset contains information of every person associated with the organization in different capacities. However, for the business problems, we're trying to solve we look for records pertaining to only job seekers and employers.

Some of the important variables we used from this dataset are -

<i>Variable</i>	<i>Description</i>
Hired/NotHired	Binary Variable denoting whether the person is hired or not
Gender__c	Male/ Female/ Prefer not to answer
Interview_Skills__c	True / False (indicates client has been told about HHUSA interview skills assistance)
Status__c	Indicates whether client was unemployed, underemployed, active duty, etc. at time of registration
Highest_Level_of_Education_Completed__c	Highest degree received by job seeking client (completed, not to include in progress)
Resume_Complete_Duration	Time taken to complete the resume assessment

Internship__c	The role accepted by the job seeking client is an internship
Is_the_Initial_Intake_Assessment_done__c	Denotes that initial assessment during the intake has been completed
Resume_Tailoring_Tips__c	Indicates client has been told about HHUSA resume tailoring assistance
Disability_percentage_60_or_above__c	Indicates proof of disability rating if rating is above 60%
Num_Activities	Total number of activities that client participated in Hire Hero activities

2. Activities

This dataset contains all the activities / interactions that HHUSA had with the client. Phone calls, emails, and text message conversations are logged here. The below mentioned fields were used to derive the features that were used in the model.

Variable	Description
OWNERID	Unique Salesforce Identifier related to Contact Owner
ACCOUNTID	Unique Salesforce Identifier related to the account record
CALLDURATIONINSECONDS	Indicates length of call (in seconds)
CORRESPONDENCE_TYPE__C	Indicates means of contact, including call, email, text message, etc.

3. Hire Information

This dataset contains information of Contacts that are hired. Hire Heroes USA migrated their Hire Data into a new object to account for the clients who were returning for additional services and finding jobs because of those services. Some of the important variables we used

from this dataset are -

Variable	Description
Job_Function_Hired_In	Job function that the client was hired into, based loosely on federal job function classifications
Salary_Range	Attained salary
Hired_With_EO_Assistance	True / False (indicates client was hired with help from internal program including job matches and interview referrals)

Since Contacts datafile has the demographic information of the clients, and Hire Information contains details about clients getting hired. We wanted to combine these two files to see how demographic information could in association with hire information, determine how a client can be hired. Using dplyr package in R, we combined these two datasets using inner join.

Derived Variables:

Below given are the variables derived from the three most critical datasets that were selected

Variable	Data Type	Description
Hired/Not Hired	Binary	If date confirmed hired is not null, we assume the client to be hired
Initial Assessment complete duration	Numeric (in minutes)	Time taken to complete the initial assessment
Number of Activities	Numeric	Total Number of interactions that HHUSA had with client
Days to Hire	Numeric (in Days)	Number of days taken for a client to get hired after initial assessment completion

List of final fields and their description (appendix)

Completeness of Data (some of the critical fields)

Column Name	Number of Missing Values	% Missing Values	Categorical Y/N	Levels
Race_c	59712	78.29	Y	7
Num_Activities	46611	61.11	N	NA
Mileage_Willing_To_Commute_c	44309	58.10	Y	6
Gender_c	23311	30.56	Y	2
Highest_Level_of_Education_Completed_c	8907	11.68	Y	8
Status_c	6628	8.69	Y	9
Used_Volunteer_Services_c	2126	2.79	Y	2
Used_Federal_Services_c	2126	2.79	Y	2
Photo_on_File_c	2070	2.71	Y	2
Resume_Tailoring_Tips_c	35	0.05	Y	2
Disability_percentage_60_or_above_c	35	0.05	Y	2
MyTrak_Employed_outside_military_c	34	0.04	Y	2
MyTrak_Federal_Resume_Review_c	34	0.04	Y	2
MyTrak_VTS_Assigned_c	34	0.04	Y	2
O2O_Initial_Assessment_Complete_c	33	0.04	Y	2
On_Job_Board_c	33	0.04	Y	2
Updated_Resume_Complete_c	33	0.04	Y	2
O2O_Hire_c	23	0.03	Y	2
MyTrak_Past_Jobs_c	1	0.00	Y	7
Hired/NotHired	0	0.00	Y	2
Interview_Skills_c	0	0.00	Y	2
Resume_Complete_Duration	0	0.00	N	NA
O2O_Program_Participant_c	0	0.00	Y	2
Internship_c	0	0.00	Y	2
Foreign_Service_c	0	0.00	Y	2
Initial_Assessment_Complete_Duration	0	0.00	N	NA
Responsive_c	0	0.00	Y	2
Volunteer_c	0	0.00	Y	2
Program_Enrollments_c	0	0.00	Y	10
HHUSA_Workshop_Participant_c	0	0.00	Y	2
Willing_to_relo_with_no_assistance_c	0	0.00	Y	2
Created_LinkedIn_account_c	0	0.00	Y	2
Is_the_Initial_Intake_Assessment_done_c	0	0.00	Y	2
Reserves_National_Guard_c	0	0.00	Y	2
Enrolled_in_School_c	0	0.00	Y	2
Finalized_HHUSA_revised_resume_on_file_c	0	0.00	Y	2
Documents_Received_c	0	0.00	Y	2
Willing_to_Relocate_c	0	0.00	Y	2

Race and Gender

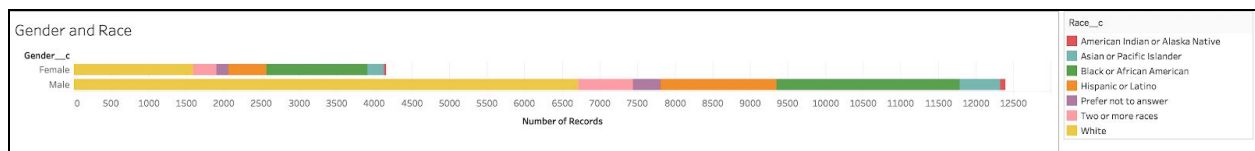


Figure 1: Initial Data exploration for Gender and Race present in the data

Salary Range

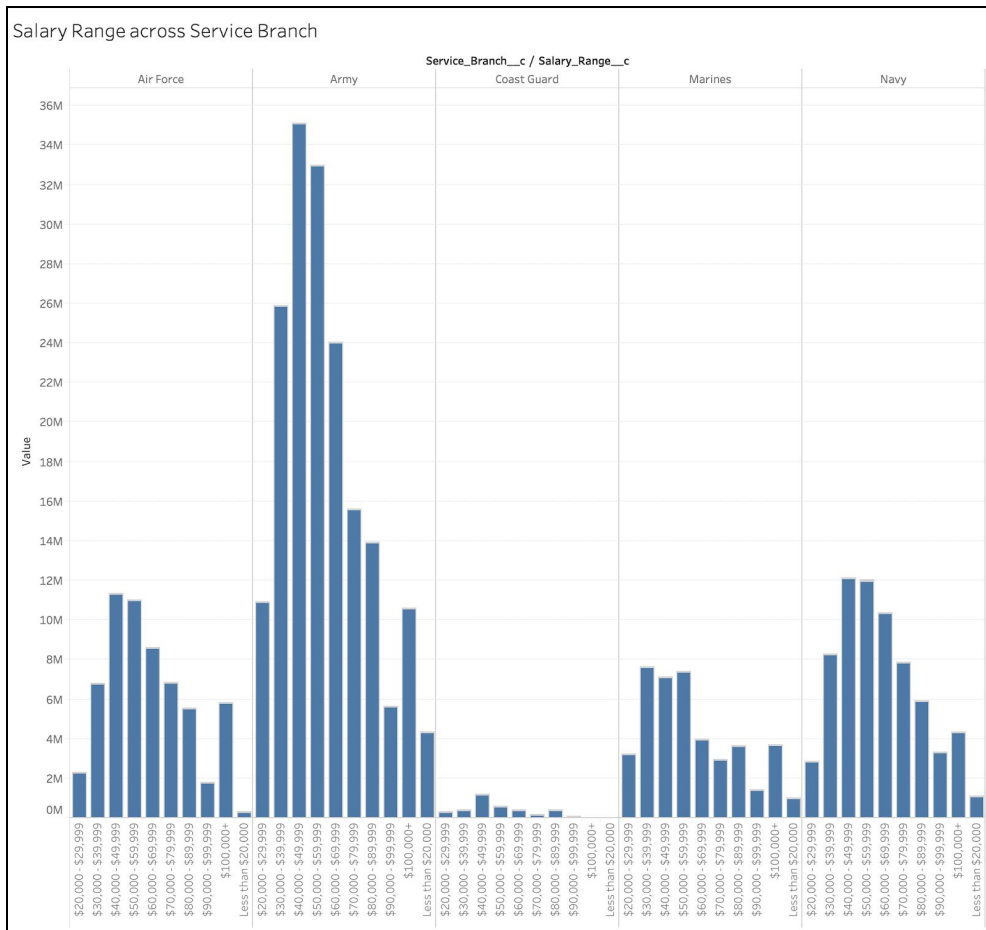


Figure 2: Salary range across service branch

Employment Status

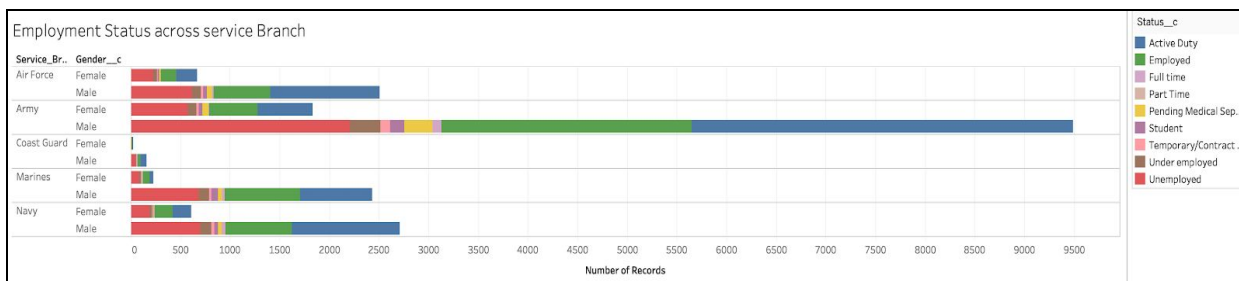


Figure 3: Employment status of service branch across gender

Disability Rating

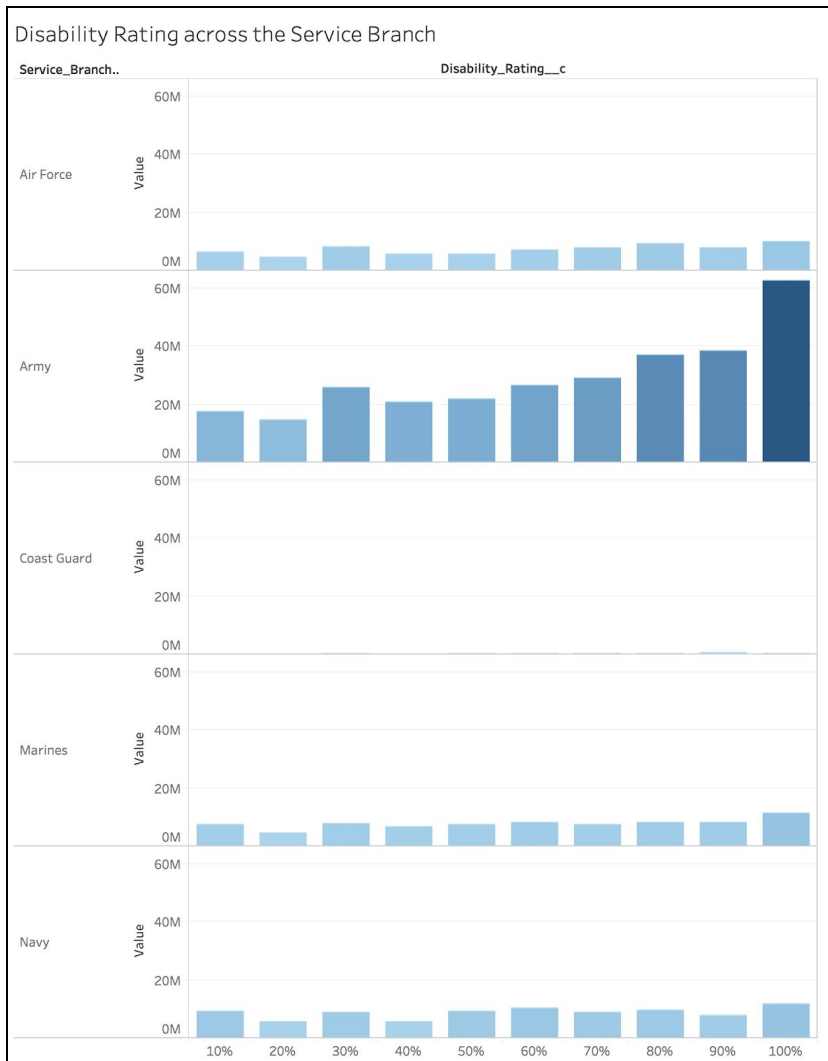


Figure 4: Disability rating across service branch

Activities

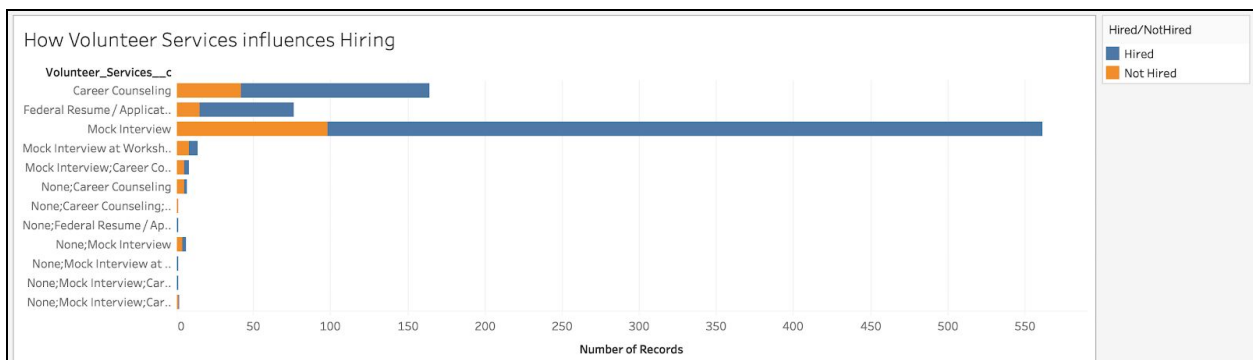


Figure 5: How Volunteering service affects people getting hired

Method

The entire data for the Teradata Data Challenge is provided by a non-profit partner, Hire Heroes USA, whose mission is to empower U.S. military members, veterans and military spouses to succeed in the civilian workforce. Hire Heroes USA uses Salesforce as their CRM, so most of their data is structured based on Salesforce's use of Objects. The data provided includes 13 spreadsheets, which we'll be using to answer the following business questions:

- ❖ Is there any relationship between the amount of time spent working with individual clients (time to complete an assessment, time to complete resume, # of logged activities, etc.) and how quickly they are employed?

To answer our research question, we first try to solve what services that are offered by Hire Heroes are pertinent to this problem along with some latent qualities that the individual has. We started with accessing what are qualities in an individual which assist him in some way, to get hired quickly (within 6 months of joining Hire Heroes)

Out of all the features given to us, we focused on 35 primary features (- original features and - derived features) that we computed after further delving into the data and with help from Hire Heroes. Originally, we had over 300+ features, with 45% of missing data.

For phase 1 of the project, we spent most of our time in data wrangling and feature selection, further reading on which can be found in the *Data Section* of this report. For phase 2 checkpoint of this project, we tried to implement a classifier using different classical and ensemble machine learning algorithms for better understanding of the features used and to get further insights on what facilitates a client to get hired quickly. In the end, we performed Survival Analysis on the top 3 features to understand the survival rate of Hired / Non Hired population with these features in context.

Phase 1:

For this phase of the project, we spent most of the time in data wrangling and feature selection. To get some idea, we ran 2 models after creating the target variable 'Hired/NotHired', which is a derived one. Before any kind of data wrangling, we used 2 models namely, Random Forest and Logistic Regression, taking only the observations with no missing variable and around 300 predictors, we got an 57% and 55% respectively, which was just above random. Then we proceeded to Data Wrangling and Feature Selection - which is covered in the *Data Section* of this report.

Phase 2:

To answer the Classification problem (i.e. if a person is hired in 6 months - 180 days or not), we used the derived column 'Hired/NotHired' as our target variable and -feature list-, and used the following predictive algorithms:

Naive Bayes Classifier ~ baseline:

This is a logic based technique which is simple yet so powerful that it is often known to outperform complex algorithms for very large datasets. This algorithm is based on Bayes theorem but with strong assumptions regarding independence. This was used as our baseline model.

For our model, we apply a Naïve Bayes model with *10-fold cross validation*, which gets 68% accuracy.

Neural Networks:

Neural Network (or Artificial Neural Network) has the ability to learn by examples. ANN is an information processing model inspired by the biological neuron system. It is composed of a large number of highly interconnected processing elements known as the neuron to solve problems. It follows the non-linear path and process information in parallel throughout the nodes. A neural network is a complex adaptive system. Adaptive means it has the ability to change its internal structure by adjusting weights of inputs.

NN Model 1

- Parameters / Tuning - 3 hidden nodes (a 64-3-1 network with 199 weights), max iteration - 100
- AUC - 0.8991

NN Model 2

- Parameters / Tuning - 10 hidden nodes (a 64-10-1 network with 661 weights), max iteration - 100
- AUC - 0.924

GLM Net:

This stands for Lasso and Elastic-Net Regularized Generalized Linear Models. It's extremely efficient procedures for fitting the entire lasso or elastic-net regularization path for linear regression, logistic and multinomial regression models, Poisson regression and the Cox model.

GLM Model

- Parameters / Tuning - 3-fold Cross Validation with default parameters
- AUC - 0.9101
- Accuracy 0.8306

GBM:

This stands for Generalized Boosted Regression Models. This is an implementation of extensions to Freund and Schapire's AdaBoost algorithm and Friedman's gradient boosting machine. Includes regression methods for least squares, absolute loss, t-distribution loss, quantile regression, logistic, multinomial logistic, Poisson, Cox proportional hazards partial likelihood, AdaBoost exponential loss, Huberized hinge loss, and Learning to Rank measures (LambdaMart).

GBM Model

- Parameters / Tuning - 3-fold Cross Validation with default parameters
- AUC : 0.9271
- Accuracy : 0.8597

Random Forest:

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it's simplicity and the fact that it can be used for both classification and regression tasks. In this post, you are going to learn, how the random forest algorithm works and several other important things about it.

For our analysis, we've used the following variants of the model:

- a) Random Forest with no tuning (default values) : Reported AUC is 0.9195
- b) Random Forest with fine tuning :

Used function tuneRF(). Parameters used:

- ntrees : 10 to 150, best value : 120
- mtry : 3 to 5, best value: 5

c) Random Forest with parameter tuning and Cross Validation:

Used 10-fold Repeated Cross-Validation, repeating 3 times. Used 'caret' package, method='rf'.

Parameters used:

- mtry : 3 to 7, best value: 7,
- num.tree: 10 to 150, best value: 120

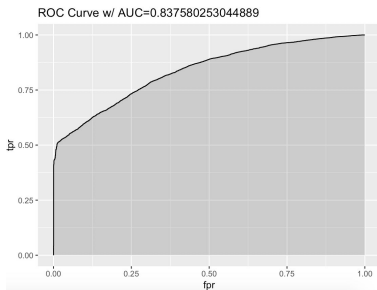
d) Random Forest with Hyperparameter tuning and Cross Validation ~ **final model**:

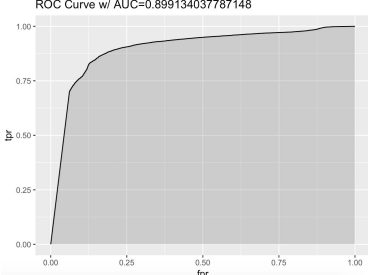
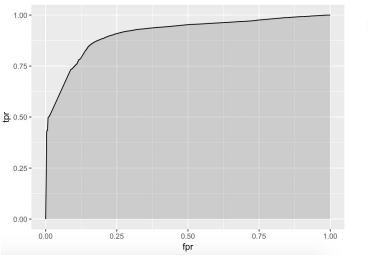
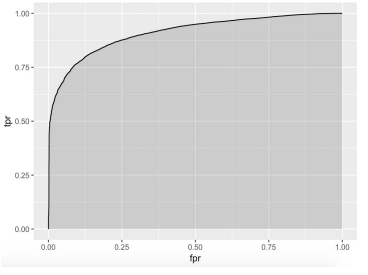
Used 10-fold Repeated Cross-Validation, repeating 3 times. Used 'caret' package, method='ranger'.

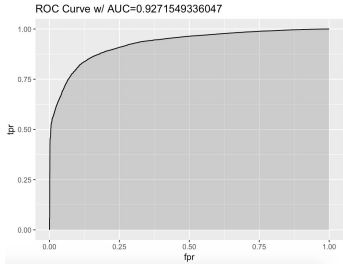
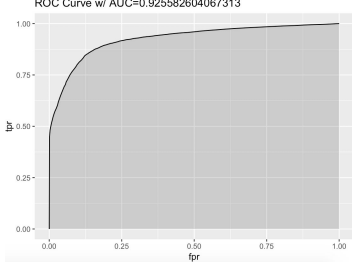
Hyperparameter used:

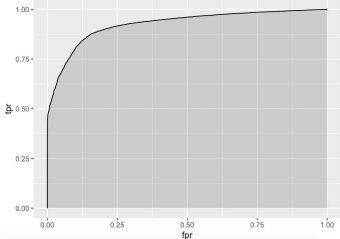
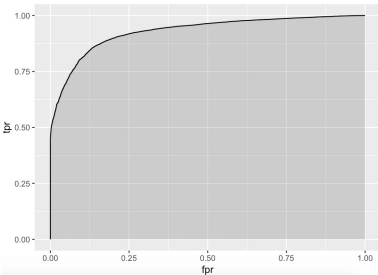
- mtry : 3 to 10, best value: 7
- splitrule: ("gini", "extratrees") - used value "gini"
- min.node.size : (1, 3, 5) - best value: 1

TABLE I: PERFORMANCE OF CLASSIFIERS

Models (w/ Parameters and Hyperparameters)	Accuracy and Area under the Curve	ROC												
No data preprocessing/feature selection, Default parameters, Predictors Count : Around 400														
Random Forest	AUC : 0.65 Accuracy : 57%													
Logistic Regression	AUC : 0.64 Accuracy : 55%													
Data Wrangling & Feature Selection, Predictors Count : Around 32														
Naive Bayes (10-fold Cross Validation) - <i>baseline</i>	Accuracy : 0.683 AUC : 0.837	<p>Confusion Matrix and Statistics</p> <table> <tr> <td></td><td colspan="2">Reference</td></tr> <tr> <td>Prediction</td><td>Hired</td><td>Not Hired</td></tr> <tr> <td>Hired</td><td>1</td><td>0</td></tr> <tr> <td>Not Hired</td><td>4836</td><td>10417</td></tr> </table> <p> Accuracy : 0.683 95% CI : (0.6755, 0.6903) No Information Rate : 0.6829 P-Value [Acc > NIR] : 0.497 Kappa : 3e-04 McNemar's Test P-Value : <2e-16 Sensitivity : 2.067e-04 Specificity : 1.000e+00 Pos Pred Value : 1.000e+00 Neg Pred Value : 6.829e-01 Prevalence : 3.171e-01 Detection Rate : 6.556e-05 Detection Prevalence : 6.556e-05 Balanced Accuracy : 5.001e-01 'Positive' Class : Hired </p> 		Reference		Prediction	Hired	Not Hired	Hired	1	0	Not Hired	4836	10417
	Reference													
Prediction	Hired	Not Hired												
Hired	1	0												
Not Hired	4836	10417												
Neural Networks	AUC : 0.8991													

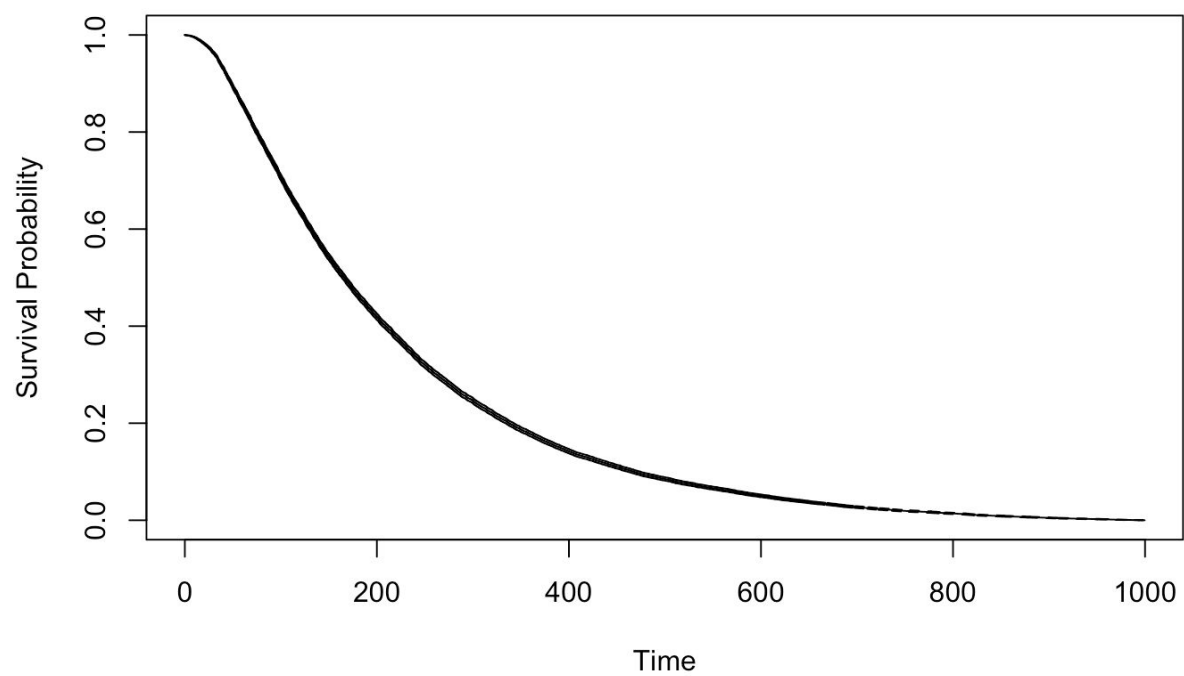
(size=3, maxit=1000)		<p>Confusion Matrix and Statistics</p> <table> <tr> <td></td><th colspan="2">Reference</th></tr> <tr> <th>Prediction</th><th>Hired</th><th>Not Hired</th></tr> <tr> <th>Hired</th><td>5751</td><td>1650</td></tr> <tr> <th>Not Hired</th><td>1521</td><td>13959</td></tr> </table> <p>Accuracy : 0.8614 95% CI : (0.8569, 0.8659) No Information Rate : 0.6822 P-Value [Acc > NIR] : < 2e-16</p> <p>Kappa : 0.6819</p> <p>McNemar's Test P-Value : 0.02302</p> <p>Sensitivity : 0.8943 Specificity : 0.7908 Pos Pred Value : 0.9017 Neg Pred Value : 0.7771 Prevalence : 0.6822 Detection Rate : 0.6101 Detection Prevalence : 0.6765 Balanced Accuracy : 0.8426</p> <p>'Positive' Class : Not Hired</p> 		Reference		Prediction	Hired	Not Hired	Hired	5751	1650	Not Hired	1521	13959
	Reference													
Prediction	Hired	Not Hired												
Hired	5751	1650												
Not Hired	1521	13959												
Neural Networks (size=10, maxit=1000)	Accuracy : 0.8587, AUC : 0.910	<p>Confusion Matrix and Statistics</p> <table> <tr> <td></td><th colspan="2">Reference</th></tr> <tr> <th>Prediction</th><th>Hired</th><th>Not Hired</th></tr> <tr> <th>Hired</th><td>5698</td><td>1658</td></tr> <tr> <th>Not Hired</th><td>1574</td><td>13951</td></tr> </table> <p>Accuracy : 0.8587 95% CI : (0.8542, 0.8632) No Information Rate : 0.6822 P-Value [Acc > NIR] : <2e-16</p> <p>Kappa : 0.6752</p> <p>McNemar's Test P-Value : 0.1443</p> <p>Sensitivity : 0.8938 Specificity : 0.7836 Pos Pred Value : 0.8986 Neg Pred Value : 0.7746 Prevalence : 0.6822 Detection Rate : 0.6097 Detection Prevalence : 0.6785 Balanced Accuracy : 0.8387</p> <p>'Positive' Class : Not Hired</p> 		Reference		Prediction	Hired	Not Hired	Hired	5698	1658	Not Hired	1574	13951
	Reference													
Prediction	Hired	Not Hired												
Hired	5698	1658												
Not Hired	1574	13951												
GLM Net (3-fold Cross Validation, returnResamp='none', summaryFunction = twoClassSummary)	Accuracy : 0.8306, AUC : 0.9101	<p>Confusion Matrix and Statistics</p> <table> <tr> <td></td><th colspan="2">Reference</th></tr> <tr> <th>Prediction</th><th>Hired</th><th>Not Hired</th></tr> <tr> <th>Hired</th><td>6035</td><td>2638</td></tr> <tr> <th>Not Hired</th><td>1237</td><td>12971</td></tr> </table> <p>Accuracy : 0.8306 95% CI : (0.8257, 0.8355) No Information Rate : 0.6822 P-Value [Acc > NIR] : < 2.2e-16</p> <p>Kappa : 0.6286</p> <p>McNemar's Test P-Value : < 2.2e-16</p> <p>Sensitivity : 0.8299 Specificity : 0.8310 Pos Pred Value : 0.6958 Neg Pred Value : 0.9129 Prevalence : 0.3178 Detection Rate : 0.2638 Detection Prevalence : 0.3790 Balanced Accuracy : 0.8304</p> <p>'Positive' Class : Hired</p> 		Reference		Prediction	Hired	Not Hired	Hired	6035	2638	Not Hired	1237	12971
	Reference													
Prediction	Hired	Not Hired												
Hired	6035	2638												
Not Hired	1237	12971												

<div>GBM (3-fold Cross Validation)</div>	<div>Accuracy : 0.8597 AUC : 0.9271</div>	<div>Confusion Matrix and Statistics</div> <div>Prediction Reference Hired Not Hired Hired 5704 1642 Not Hired 1568 13967</div> <div>Accuracy : 0.8597 95% CI : (0.8551, 0.8642) No Information Rate : 0.6822 P-Value [Acc > NIR] : <2e-16</div> <div>Kappa : 0.6773</div> <div>McNemar's Test P-Value : 0.1976</div> <div>Sensitivity : 0.7844 Specificity : 0.8948 Pos Pred Value : 0.7765 Neg Pred Value : 0.8991 Prevalence : 0.3178 Detection Rate : 0.2493 Detection Prevalence : 0.3211 Balanced Accuracy : 0.8396</div> <div>'Positive' Class : Hired</div> <div></div>
<div>Fine-tuned RF (120 trees, 5 nodes)</div>	<div>AUC : 0.9255</div>	<div>Prediction Reference Hired Not Hired Hired 5753 1527 Not Hired 1519 14082</div> <div>Accuracy : 0.8669 95% CI : (0.8624, 0.8713) No Information Rate : 0.6822 P-Value [Acc > NIR] : <2e-16</div> <div>Kappa : 0.6931</div> <div>McNemar's Test P-Value : 0.8991</div> <div>Sensitivity : 0.7911 Specificity : 0.9022 Pos Pred Value : 0.7902 Neg Pred Value : 0.9026 Prevalence : 0.3178 Detection Rate : 0.2514 Detection Prevalence : 0.3182 Balanced Accuracy : 0.8466</div> <div>'Positive' Class : Hired</div> <div></div>

10 fold Repeated Cross-Validation, mtry =7, num.tree = 120	AUC : 0.9253	<div>Confusion Matrix and Statistics</div> <div> Reference Prediction Hired Not Hired Hired 5710 1497 Not Hired 1562 14112 </div> <div> Accuracy : 0.8663 95% CI : (0.8618, 0.8707) No Information Rate : 0.6822 P-Value [Acc > NIR] : <2e-16 Kappa : 0.6909 McNemar's Test P-Value : 0.2472 Sensitivity : 0.7852 Specificity : 0.9041 Pos Pred Value : 0.7923 Neg Pred Value : 0.9003 Prevalence : 0.3178 Detection Rate : 0.2496 Detection Prevalence : 0.3150 Balanced Accuracy : 0.8446 'Positive' Class : Hired </div> <div> ROC Curve w/ AUC=0.925350141603308  </div>
10- fold Repeated Cross-Validation (repeats 3 times) , splitrule = gini and min.node.size = 3, metric = ROC, mtry =8 (using Tune grid)	AUC : 0.9290 Accuracy: 0.8675	<div>Confusion Matrix and Statistics</div> <div> Reference Prediction Hired Not Hired Hired 5713 1473 Not Hired 1559 14136 </div> <div> Accuracy : 0.8675 95% CI : (0.863, 0.8719) No Information Rate : 0.6822 P-Value [Acc > NIR] : <2e-16 Kappa : 0.6934 McNemar's Test P-Value : 0.1227 Sensitivity : 0.7856 Specificity : 0.9056 Pos Pred Value : 0.7950 Neg Pred Value : 0.9007 Prevalence : 0.3178 Detection Rate : 0.2497 Detection Prevalence : 0.3141 Balanced Accuracy : 0.8456 'Positive' Class : Hired </div> <div> ROC Curve w/ AUC=0.929080557809127  </div>

Survival Analysis: In Random Forest classification model, using the features from the dataset we predicted the binary outcome of Hired & Not Hired veterans. But it is also critical to determine those features that affect the time it takes for an individual to get hired. We use survival analysis to determine this effect. Survival analysis is used to analyze data in which the time until the event is of interest. The response is often referred to as a failure time(not hired) or event time(hired). In survival analysis we are identifying the difference in the rate at which a client is getting hired within 'n' days for the different variables under consideration.

Graph: Overall survival Probability



1.4 Results & Discussions

Based on the above Table I, we can clearly see that the highest AUC is 92.9% and the lowest is 83.7%, along with the highest accuracy of 86.75%, and the lowest of 68.3%. This belongs to the Random Forest classifier with 10-fold cross validation, 120 trees and 5 variables for splitting at each tree node. An average of 19,620 instances out of 22,881 instances is found to be classified correctly with the highest score of 19,849 instances and lowest of 19,006 instances. Here our focus is on achieving the maximum True Positives, as we want to be able to predict how many people have been hired correctly. As this is a classification problem we choose the evaluation criteria to be area under the curve and our correct instance classifications instead of an accuracy measure.

Kappa statistic is a measure used to test interrater reliability, to distinguish between the collected data and its validity. Based on the kappa statistic criteria, the optimal model using Random Forest has a score of 0.69, which proves to be substantial for the accuracy of this classification problem(Mary L. McHugh, Oct 2012). On average the Kappa statistic for our models was 0.67 ranging from highest score of 0.69 and lowest score of 0.62. This proves that the resulting models all performed substantially well in the interrater reliability.

The results obtained when compared to the paper studied by (A. K. Pal, and S. Pal, Nov 2013) actually prove to be doing a better job. The result from the study provided an accuracy of 86.15% while our results provide an accuracy of 86.75%. In spite of the accuracy being just slightly better, we can infer that our model is significant and possibly even better to the studies conducted in the related field. This is probably because we explored more complex algorithms while the paper only studies basic to medium algorithms.

Using our final Random Forest model with an AUC of 92.9%, Boruta and Logistic Regression results, we can generate the two tables namely Table II and Table III shown in appendix, where Table II represents the topmost important features list with its impact and Table III shows the impact of demographic features on our classification. We use these to help us determine how to get a veteran hired quickly.

Based on Table II from appendix we observe that the most important features, which can be categorized into Active Participation, Specialist Consulting and Training Workshops. This is a significant finding as we can clearly see a direct correlation with the research question of determining if there is a relationship between the time spent with clients to how quickly they are hired. Spending time with clients here mean supporting them through trainings, providing consultation, and getting them to actively participate in various activities to stay on top of the industry standards like for resumes. We observe that the clients who have been actively participating in various activities, received consulting from specialists and have attended training workshops are more likely to be hired than those who did not.

Table III from appendix shows us some interesting findings indicating that gender and disability play an important role in determining if a client is going to be hired in less than 180 days or not. We see that gender = male has a greater positive impact in getting a person hired in

comparison to a person with gender = female. Similarly, a person with a disability of greater than 60% is less likely to be hired than a person without a disability. We can say that gender male consists of the majority class so the result from this would be inconclusive unless the data is balanced for male and females. Similarly, the disability greater than 60% is only taken into account after ignoring the majority Null class or interpreted as 'Prefer not to answer' class. These, even if significant in our logistic regression, did not prove to be amongst our topmost important features. Therefore, we need more evidence to be able to prove their impact, but our results do raise a flag that this is something that needs to be explored further to validate the findings observed.

We also found some insights from our survival analysis which definitely helps us answer our problem statement partly. The main takeaway from this analysis is that if a client was assigned a specialist who helped them through the transition process, then the chances of the client getting hired in less than 180 days in the civilian market was significantly higher. This factors into our recommendations as well and therefore, definitely is an important finding from a result point of view.

1.5 Conclusion

After going through the entire process of cleaning and understanding our data and interpreting the results, we have concluded there are a few significant factors which influence a veteran's employability in the civilian market. As a key highlight of our entire process, we explored the dataset provided by HHUSA to explore the relationship between the amount of time spent working with clients, their demographics and how quickly they are employed in the civilian job markets. On that basis, we built multiple classification algorithms and found the optimal result using Random Forest that predicts if a person will be hired in less than 180 days with an AUC of 92.9%.

With all these, we concluded that the most important features can be categorized into Active Participation, Specialist Consulting and Training Workshops and, these are the factors to be considered by Hire Heroes while looking at strategizing and understanding their client requirements better. Certain demographic features like Gender, Disability and Highest Level of Education show us some correlation but we need more supporting evidence to be able to validate this. Our current data suggests that these features could be skewed and therefore this could be a potential caveat to our model. However, ignoring these features could also result in potential decision-making errors and some extra validation techniques may be required to conclude with higher confidence.

In response to our initial research question, we feel that we have successfully answered it and have been able to provide them with recommendations that could help them secure more funding to improve the quality of service currently provided. A few pointers that give a very high level understanding of our recommendations that we feel Hire Heroes should adopt are assigning transition specialists to assist clients in their tasks like resume tailoring, and hosting mock interviews which is the top request from clients, encourage clients to attend workshops conducted by HHUSA and encourage clients to be actively conducting activities like working on resume in order to stay on top of the industry standards. Lastly, more onsite

opportunities could work to their advantage in comparison to virtually offered sessions. We feel all these recommendations could help a veteran get hired quickly in the civilian job market.

However, as a caveat to our model, some high-level shortcomings would be that currently there is no framework to increase the correctness and completeness of data. In addition, certain critical fields such as Rank, Employed Position, Skillset of hired veterans must be given importance and must be used to analyze and aid the unemployed veterans.

In the end, the most advantageous move for Hire Heroes would be, to leverage these recommendations and harness the power of these results to improve quality of talent hired, increase efficiency (quick reach to civilian market) and quality of the service they provide to their clients.

Table II: Topmost Important Features and their Impact

<i>Feature</i>	<i>Impact Description</i>
Number of Activities	The larger the number of activities performed by a client, the greater the chance of being hired
MyTrack VTS Assigned	A specialist (specialist consultant) assigned, increases the chance of a client being hired than when a specialist is not assigned
Finalized Revised Resume on File	If a final resume is on file, then probability of getting hired increases than when a resume is not on file
Initial Assessment Complete Duration	The longer the duration in completing the initial assessment, the lower the chances of being hired
Resume Tailoring Tips	If client did receive resume tailoring tips, then there is a higher chance of being hired than when client did not receive resume tailoring tips
HHUSA Workshop Participant	If client attended HHUSA workshop, then a greater chance of being hired than when client did not attend HHUSA workshop

Table III: Demographic Features and their Impact

<i>Feature</i>	<i>Impact</i>	<i>Description</i>
Gender	Positive	A male client has a higher probability of getting hired than a female client
Race	Not Significant	Cannot Interpret
Highest Level of Education	Positive	A client with a master's degree has a higher probability of getting hired than a client without any degree
Disability > 60%	Negative	A person with a disability of >60% is less likely to be

		hired than a person without a disability
--	--	--

Table IV: List of final fields and their description

<i>Variable</i>	<i>Description</i>
Hired/NotHired	Binary Variable denoting whether the person is hired or not
Gender__c	Male/ Female/ Prefer not to answer
Interview_Skills__c	True / False (indicates client has been told about HHUSA interview skills assistance)
Status__c	Indicates whether client was unemployed, underemployed, active duty, etc. at time of registration
Highest_Level_of_Education_Completed__c	Highest degree received by job seeking client (completed, not to include in progress)
Resume_Complete_Duration	Time taken to complete the resume assessment
O2O_Program_Participant__c	True/False (indicates job seeking client is also pursuing certification through the O2O program)
O2O_Initial_Assessment_Complete__c	True / False (indicates O2O coordinator has had first contact with client)
MyTrak_Employed_outside_military__c	Client has indicated they are / are not employed outside of the military (MyTrak response)
Internship__c	The role accepted by the job seeking client is an internship
Foreign_Service__c	Indicates job seeking client has served in the military in a location other than the United States
Initial_Assessment_Complete_Duration	Time taken to complete initial assessment
Responsive__c	True / False (Indicates whether candidate /

	employer / donor has responded to outreach)
Volunteer__c	Binary (if 1; Contact is Volunteer)
On_Job_Board__c	True / False (indicates client has created a profile on HHUSA job board)
Program_Enrollments__c	Number of programs, client has enrolled for
Updated_Resume_Complete__c	Typically used for returning clients further revisions to client resume beyond initial HHUSA resume review / development
HHUSA_Workshop_Participant__c	True / False (indicates job seeking client has attended a workshop hosted by HHUSA)
Willing_to_relo_with_no_assistance__c	Client is willing to move to a new location for a job with no financial assistance
Created_LinkedIn_account__c	Binary variable (1 denotes; Client has created a LinkedIn account)
MyTrak_Federal_Resume_Review__c	Client submitted a request within MyTrak to have their federal resume reviewed
MyTrak_VTS_Assigned__c	Client has been connected to a Transition Specialist
MyTrak_Past_Jobs__c	Client has filled in information regarding any previous roles (Past Jobs section in MyTrak)
Photo_on_File__c	Photo of the contact is available for use
Is_the_Initial_Intake_Assessment_done__c	Denotes that initial assessment during the intake has been completed
Resume_Tailoring_Tips__c	Indicates client has been told about HHUSA resume tailoring assistance
Reserves_National_Guard__c	Indicates job seeking client's military service was Reserve status or National Guard membership

Enrolled_in_School__c	True / False (indicates whether the job seeking client is currently pursuing a degree or certification)
Disability_percentage_60_or_above__c	Indicates proof of disability rating if rating is above 60%
Finalized_HHUSA_revised_resume_on_file__c	True / False (indicates new resume created / revised)
Documents_Received__c	True / False (indicates whether requested documents including proof of service and original resume have been submitted by the client)
Willing_to_Relocate__c	Job seeking client is willing to move to a new location (other than current residence)
Used_Volunteer_Services__c	Participated in a mentoring session or mock phone interview
Used_Federal_Services__c	Job Seeking client had a federal resume reviewed by HHUSA team
Num_Activities	Total number of activities that client participated in Hire Hero activities

Code Here:

- 1) Teradata Code Link and a separate RMD File
- 2) Boruta.RMD submitted as separate file
- 3) MoreModels.RMD submitted as separate file
- 4) Survival Analysis Code Link and submitted a separate RMD file

3 Reference

1. **U.S. Department of Labor** Jan 2019
2. **U.S. Department of Labor** Jan 2019
3. **U.S. Department of Labor** Jan 2019
4. Currie, J., and Schwandt, H. 2014; Lin, M. W., Sandifer, R., and Stein, S. 2011
5. Hire Heroes USA Feb 2019, "Top-rated Veteran Employment Non-profit Empowers Veterans and Military Spouses to Secure Meaningful Jobs"
6. Natalie Gross, July 2018, "Study: Companies still don't understand veterans"
7. Hall, Harrell, Bicksler, Stewart, & Fisher, 2014; Harrell & Berglass, 2014, "Exploring U.S. Veterans' post-service employment experiences"
8. Lisa Nagorny and Dan Pick, 2019, "5 Reasons Why Employers Are Not Hiring Vets"
9. Jesse Canella, CEO – Military Talent Group and David Pollard, Chairman - PredictiveHR July 2018, "Military Veteran Talent Development Under the Watchful Eye of Predictive Analytics"
10. A. K. Pal, and S. Pal, Nov 2013, "Classification Model of Prediction for Placement of Students" in MECS (<http://www.mecspress.org/>) DOI: 10.5815/ijmecs.2013.11.07
11. Mary L. McHugh, Oct 2012, "Interrater reliability: the kappa statistic"