H1B Visa Final Report

Lo Yu Liu, Rohith Ramineni, Jessica Starck

April 8, 2020

Overview

Our analysis is around international work visas and the benefits to employers in predicting the outcome. The dataset contains information regarding the applications filed in the years 2015 to 2018 for applicants working in different sectors of industries in various countries. It includes details related to their employment opportunity and employer, along with application details like date of submission and associated attorney information. While the data includes visa types applicable for the US, Australia, Singapore, and Chile, the majority of the applications are filed for the visa type H1-B which corresponds to the working visa permit in the United States of America for non-immigrant temporary workers. This visa type is sponsored by a U.S citizen employer and requires a minimum annual salary of $60,000 (as of 2019) for the applicant. Even then, there are some other important conditions to be satisfied to obtain the H1-B visa. Some of the important variables are given in the dataset along with their visa statuses. The four different outcomes in the visa statuses are:

- **Certified:** The visa is granted
- **Denied:** The visa is rejected
- **Certified-Withdrawn:** The visa is granted but the application is withdrawn by the employer for some reason unspecified.
- **Withdrawn:** The application is withdrawn before the decision is finalized.

The objective of this predictive modeling is to create a model such that based on the given variables, the outcome of the visa status is predicted by selecting the relevant factors. This helps the client (employer) to make business decisions regarding the sponsoring of applicants for H1-Bs. For example, based on the confidence in the certification of the applicant, the employer can begin planning and preparing projects accordingly as well as proactively adjust staffing levels as needed.

Methodology

The dataset includes applications from 2015 to 2018 with around 50 variables. After exploratory analysis, it was clear that the applications, and therefore required data entry fields, have changed overtime and additional pre-processing of the data was required prior to conducting any predictive modeling. This process included removal of rows with too many missing values (i.e. no information related to that column) and columns which either were not relevant or included little variation in their values. Some variables, such as phone numbers, cities, job title, etc. had too many levels and are not practically relevant to the status of the visa. Other variables, variables that are only relevant to “completed” applications, such as DECISION_DATE were removed as they would not be available for in-progress applications. Lastly, variables like EMPLOYMENT_START_DATE, EMPLOYMENT_END_DATE, or CASE_SUBMITTED (date), are time-based variables that were transformed into useable variables applicable to future applications. For example, a new variable was created for the employment duration (in months) and the CASE_SUBMITTED date was reduced to only the submitted month to evaluate the seasonal influence. Another transformation we had to make was regarding the pay rate for the applicant. In the years 2017 and 2018, the wages were given in the form of a range, ‘WAGE_RATE_OF_PAY_FROM’ and ‘WAGE_RATE_OF_PAY_TO’. We calculated the average of these two values as the salary of the applicant and a new variable ‘WAGE_MEAN’ was created; replacing the above two variables. Lastly, the wage rates were given in hourly, weekly, biweekly, and yearly basis so they all were converted into yearly rates to standardize the unit of measure. The remaining variables and a summary of the values can be found in Figures 2 and 3 in Appendix A.

After the cleaning of the dataset is done, all the variables are converted into appropriate units such as numeric, factor, etc. and the dataset is split into training and testing subsets. The outcome variable CASE_STATUS has four outcomes as mentioned above which were converted into two results for the model to become a binary classification problem. The following are the two options considered for this conversion.

> **Option-1:** The 'Certified_Withdrawn' is considered to be in 'Certified' case status as a decision was made in favor of certifying only to be withdrawn afterwards. Case status of 'Withdrawn' is considered to be 'Denied' since the actual outcome is unknown. Hence the four outcomes of the case status are converted into two: either certified or denied.

> **Option-2:** The outcomes are classified into 'successful' or 'unsuccessful'. The 'Certified' case status is considered successful and the rest outcomes go into the unsuccessful category assuming even 'Certified-Withdrawn' ultimately result in an unfavorable outcome.

From the standpoint of the client (the employer), it is most desired to know when the outcome results in an employee that is working for the client. In further research, we found that a common reason for a 'Certified-Withdrawn' case status is as follows:

- An employee's H1-B application can be filed by utmost 5 employers at a time and if the visa is granted in more than one application, then the employee chooses to withdraw the other visa applications (i.e., from the employer where the employee will not be working).

Considering this example, we chose Option 2 from above since the "case withdrawn" status is a negative for the client and hence considered to be 'unsuccessful'. While it would be very helpful to know the reason for withdrawal, option 2 also provide a more conservative approach for our client's interests.

Five classification methods of different types were chosen to offer a variety of approaches and support for complexity in the data. The five models used are Logistic regression, PLS, Regression tree, Support Vector Machines and Random Forest classifier. Cross validation with ten folds is performed for the dataset for each of the models to increase the effectiveness of each observation in the dataset. To determine the best fit, the interpretability, ROC metric and the sensitivity and specificity values are taken into consideration.

Results

After cleaning the dataset by following the above steps and segregating the CASE_STATUS into two outcomes, the distribution between successful and unsuccessful classifications is 88% and 12%, respectively. The division between CASE_STATUS before and after re-classification can be seen in Figure 1 in Appendix A. The final variables (as seen in figure 2) that are used for generating the model along with their summary is shown in figure 3. The successful case status applications are around 8027 while the unsuccessful applications are 1011 in this training dataset, whereas the successful and unsuccessful observations in the testing dataset are 879 and 121 respectively which can be seen in figure 4. Figure 5 shows the comparison of the ROC Curves for all models. It is observed that the Area Under the Curve (AUC) values of all the models are close ranging from 68.2% to 70.9%. Hence as the area under curve is similar for all the models, interpretability and simplicity is considered for the model selection. Logistic regression model (GLMnet) is chosen as the best option to perform the classification due to its simplicity as part of the Linear Regression family compared to the more complicated non-linear (SVM) and ensemble (Random Forest) models. For Logistic Regression, the threshold is selected by considering the optimal value of sensitivity and specificity by taking the intersection of true positive rate (rate of correctly predicting the 'successful' case status in applications-"sensitivity") and true negative rate (rate of correctly predicting the 'unsuccessful' visa case status-"specificity") which can be seen in figure 6. Different values (0.5,0.85,0.90,0.95) are taken for thresholds and checked for the preferred sensitivity and specificity. Due to the client's interest in detecting 'unsuccessul' applications (negative classification in the model) we emphasized the specificity values and chose a threshold of 0.85; resulting in a sensitivity is 0.88 and the specificity is 0.40, which can be seen in Figure 7 in Appendix A.

Critical Assessment

One concern we have is that our dataset may not be exhaustive of all H1-B applicants, as it is found that only 1% of the total applications had a status of denied whereas further research from other sources has shown around 9% is the denial percentage for the H1-B visa over the years 2015 to 2019. We are also aware that the large discrepancy between the percent certified and not certified can bias the results towards the certified visa statuses. Also it is assumed that “Withdrawn” case status is considered to be negative for the client and hence is added to the unsuccessful category, but the true reason and outcome of the withdrawn application is unknown. There may be possible circumstances where an application is withdrawn because the employee has other means to support their work with the client. Reasons for the ‘certified-withdrawn’ and ‘withdrawn’ applications can help improve the efficiency of this model as the reason helps to sort the data more clearly and effectively into successful and unsuccessful case outcomes. Lastly, due to the nature of a federal program such as the H1-B visas and the inherent relationship between international countries, our predictive model will be unable to reflect any future policy or process changes relating to this program.

Appendix A

Figure 1: Distribution of CASE_STATUS Classifications
Left = Original Classifications, Right = Revised Classifications

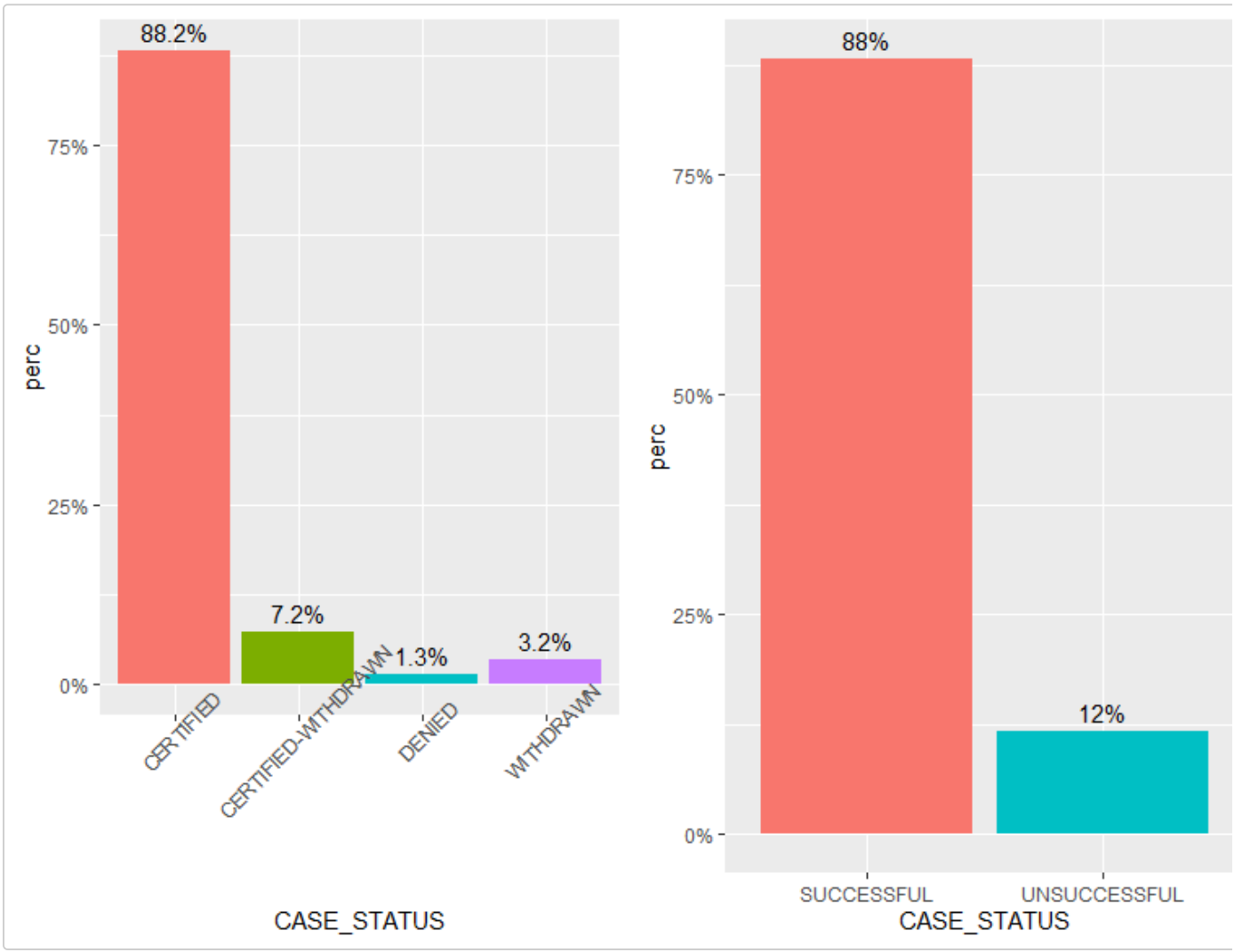


Figure 2: Final variables and data types used in models

```
## 'data.frame':    1129794 obs. of  19 variables:
##  $ CASE_STATUS               : Factor w/ 2 levels "SUCCESSFUL","UNSUCCESSFUL": 2 2 2 2 2 2
2 2 2 2 ...
##  $ CASE_SUBMITTED            : num  2 2 2 2 2 2 2 2 2 2 ...
##  $ EMPLOYER_STATE            : Factor w/ 59 levels "","AK","AL","AR",..: 50 33 40 9 7 7 7 7
7 7 ...
##  $ AGENT_ATTORNEY_STATE      : Factor w/ 57 levels "","AK","AL","AR",..: 48 38 38 1 6 6 6 6
6 6 ...
##  $ NAICS_CODE                : int  213112 523110 523110 611310 333295 333295 333295 333295
541511 333295 ...
##  $ PW_SOURCE                 : Factor w/ 8 levels "","CBA","DBA",..: 4 4 4 4 4 4 4 4 4 7
...
##  $ PW_SOURCE_YEAR            : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2016 ...
##  $ WORKSITE_STATE            : Factor w/ 59 levels "","AK","AL","AR",..: 51 33 40 8 6 6 43
6 6 6 ...
##  $ year                      : int  2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 ...
##  $ AGENT_REPRESENTING_EMPLOYER: Factor w/ 3 levels "","N","Y": 3 3 3 2 3 3 3 3 3 3 ...
##  $ NEW_EMPLOYMENT            : int  1 1 1 0 1 1 1 1 1 1 ...
##  $ CONTINUED_EMPLOYMENT      : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ CHANGE_PREVIOUS_EMPLOYMENT : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ CHANGE_EMPLOYER           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AMENDED_PETITION          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ H1B_DEPENDENT             : Factor w/ 3 levels "","N","Y": 2 2 2 2 2 2 2 2 2 2 ...
##  $ LABOR_CON_AGREE           : Factor w/ 3 levels "","N","Y": 1 3 3 3 3 3 3 3 3 3 ...
##  $ SOC_CATEGORY              : int  15 15 13 19 17 17 17 17 15 17 ...
##  $ WAGE_RATE_YEARLY          : num  84403 35000 42500 24564 70000 ...
```

Figure 3: Summary of Training Dataset Values

```
##        case_status   case_submitted   employer_state agent_attorney_state
##  SUCCESSFUL  :8027   Min.   : 1.000   CA     :1568          :3121
##  UNSUCCESSFUL:1011   1st Qu.: 3.000   TX     :1189   CA     :1149
##                      Median : 3.000   NJ     :1131   NY     : 925
##                      Mean   : 4.609   NY     : 634   TX     : 528
##                      3rd Qu.: 6.000   IL     : 577   IL     : 429
##                      Max.   :12.000   PA     : 419   MA     : 399
##                                       (Other):3520   (Other):2487
##    naics_code                                pw_source    pw_source_year
##  Min.   :  1112   OES                           :7330   Min.   :2008
##  1st Qu.:454111   Other                         :1643   1st Qu.:2016
##  Median :541511   CBA                           :  65   Median :2017
##  Mean   :448190                                 :   0   Mean   :2017
##  3rd Qu.:541511   DBA                           :   0   3rd Qu.:2017
##  Max.   :927110   OES (ACWIA - Higher Education):   0   Max.   :2018
##                   (Other)                       :   0
##  worksite_state      year      agent_representing_employer new_employment
##  CA     :1746   Min.   :2017    :   0                      Min.   :  0.0000
##  TX     : 942   1st Qu.:2017   N:2849                      1st Qu.:  0.0000
##  NY     : 799   Median :2018   Y:6189                      Median :  0.0000
##  NJ     : 596   Mean   :2018                              Mean   :  0.9054
##  IL     : 442   3rd Qu.:2018                              3rd Qu.:  1.0000
##  WA     : 371   Max.   :2018                              Max.   :100.0000
```

```
##  (Other):4142
##  continued_employment change_previous_employment change_employer
##  Min.   : 0.0000      Min.   : 0.0000            Min.   : 0.0000
##  1st Qu.: 0.0000      1st Qu.: 0.0000            1st Qu.: 0.0000
##  Median : 0.0000      Median : 0.0000            Median : 0.0000
##  Mean   : 0.3442      Mean   : 0.1202            Mean   : 0.2585
##  3rd Qu.: 0.0000      3rd Qu.: 0.0000            3rd Qu.: 0.0000
##  Max.   :50.0000      Max.   :50.0000            Max.   :50.0000
##
##  amended_petition h1b_dependent labor_con_agree  soc_category
##  Min.   : 0.000    :   0         :5511          Min.   :11.00
##  1st Qu.: 0.000   N:5802        N: 123          1st Qu.:15.00
##  Median : 0.000   Y:3236        Y:3404          Median :15.00
##  Mean   : 0.286                                 Mean   :16.12
##  3rd Qu.: 0.000                                 3rd Qu.:15.00
##  Max.   :50.000                                 Max.   :53.00
##
##  wage_rate_yearly
##  Min.   : 12282
##  1st Qu.: 35445
##  Median : 45500
##  Mean   : 58863
##  3rd Qu.: 71958
##  Max.   :648000
##
```

Figure 4: Distribution of CASE_STATUS in Test Dataset

case_status	Observations
SUCCESSFUL	879
UNSUCCESSFUL	121

Figure 5: ROC Curves for Each Model

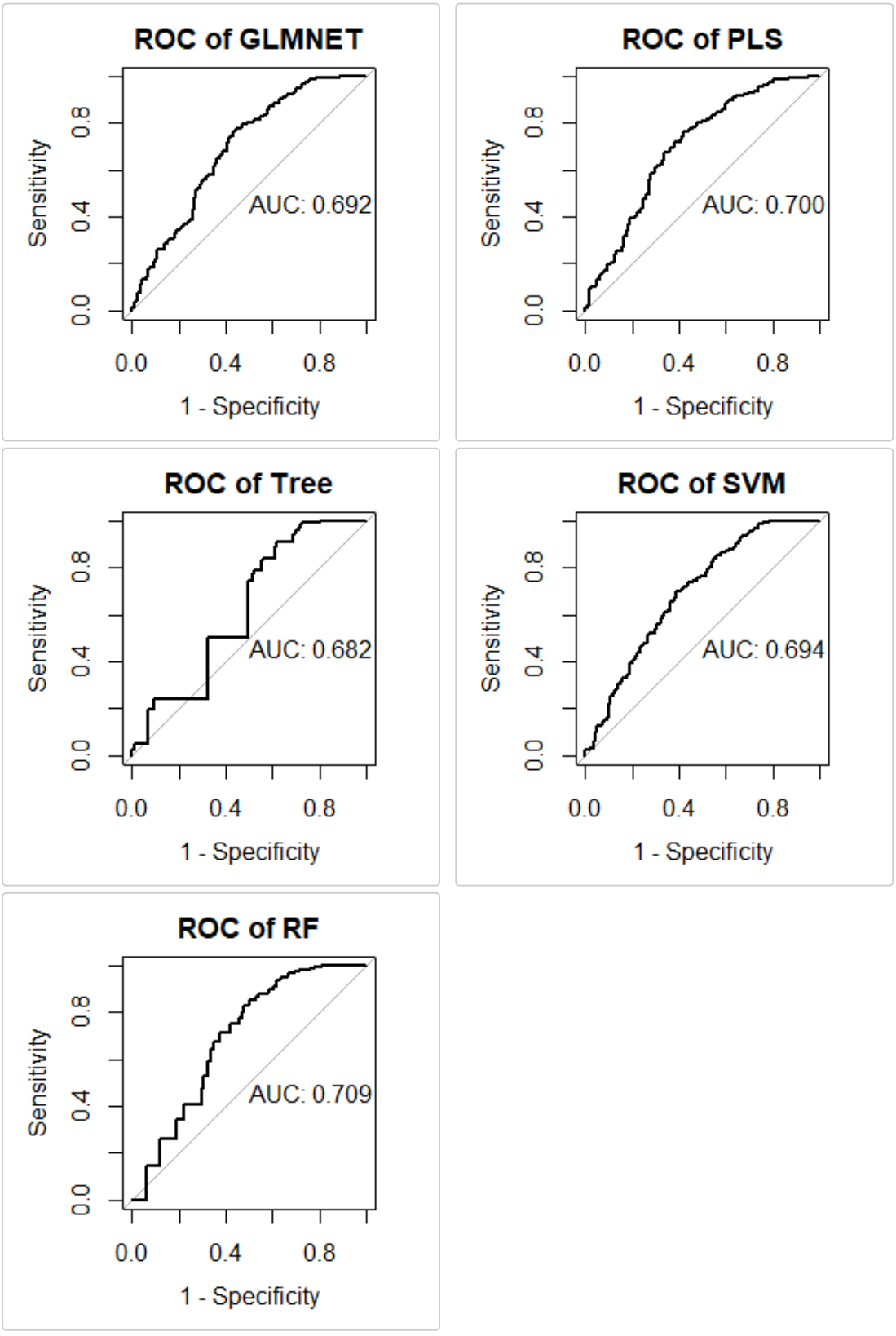


Figure 6: Sensitivities and Specificities Intersection for GLMnet
used to determine threshold for ROC curve

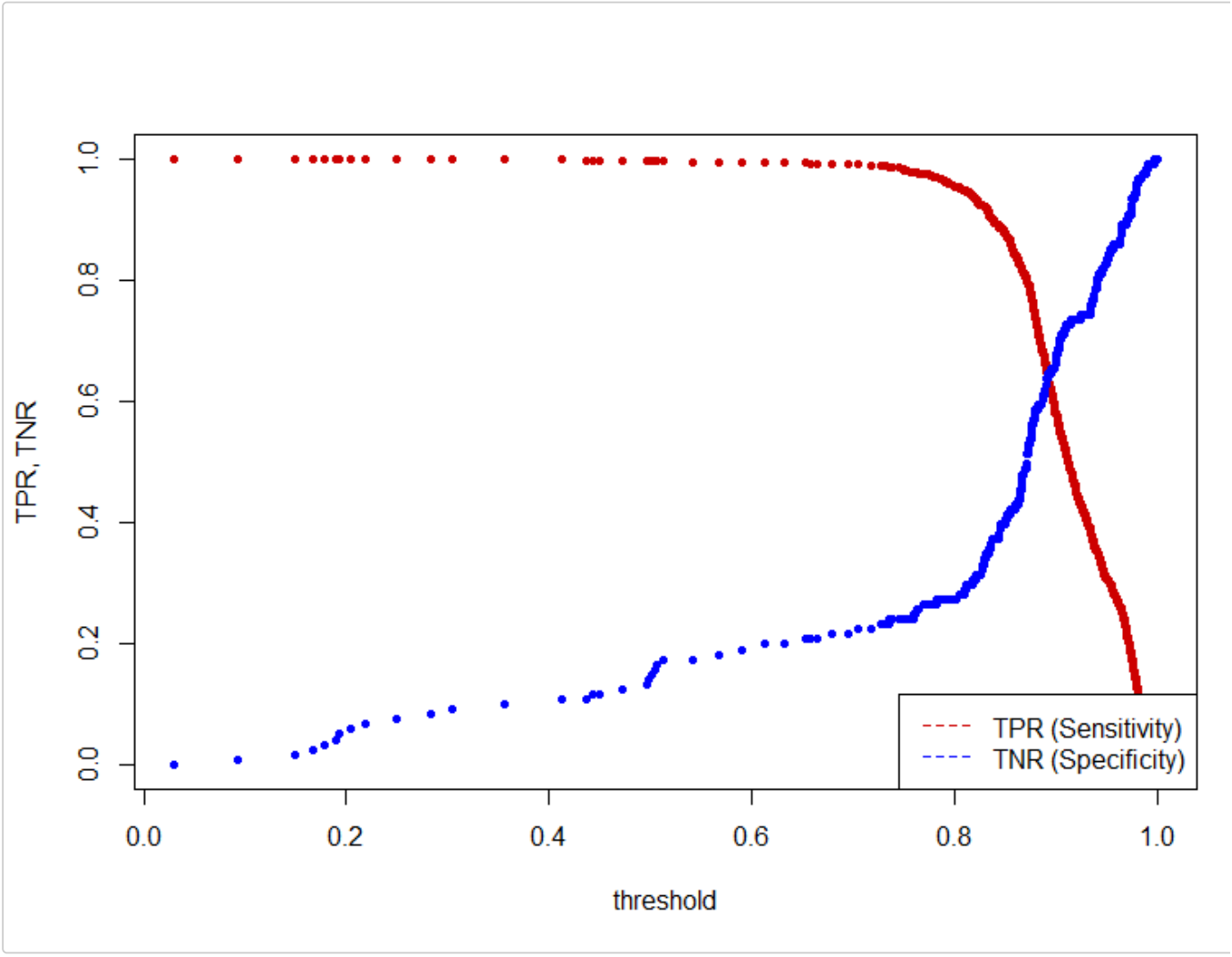
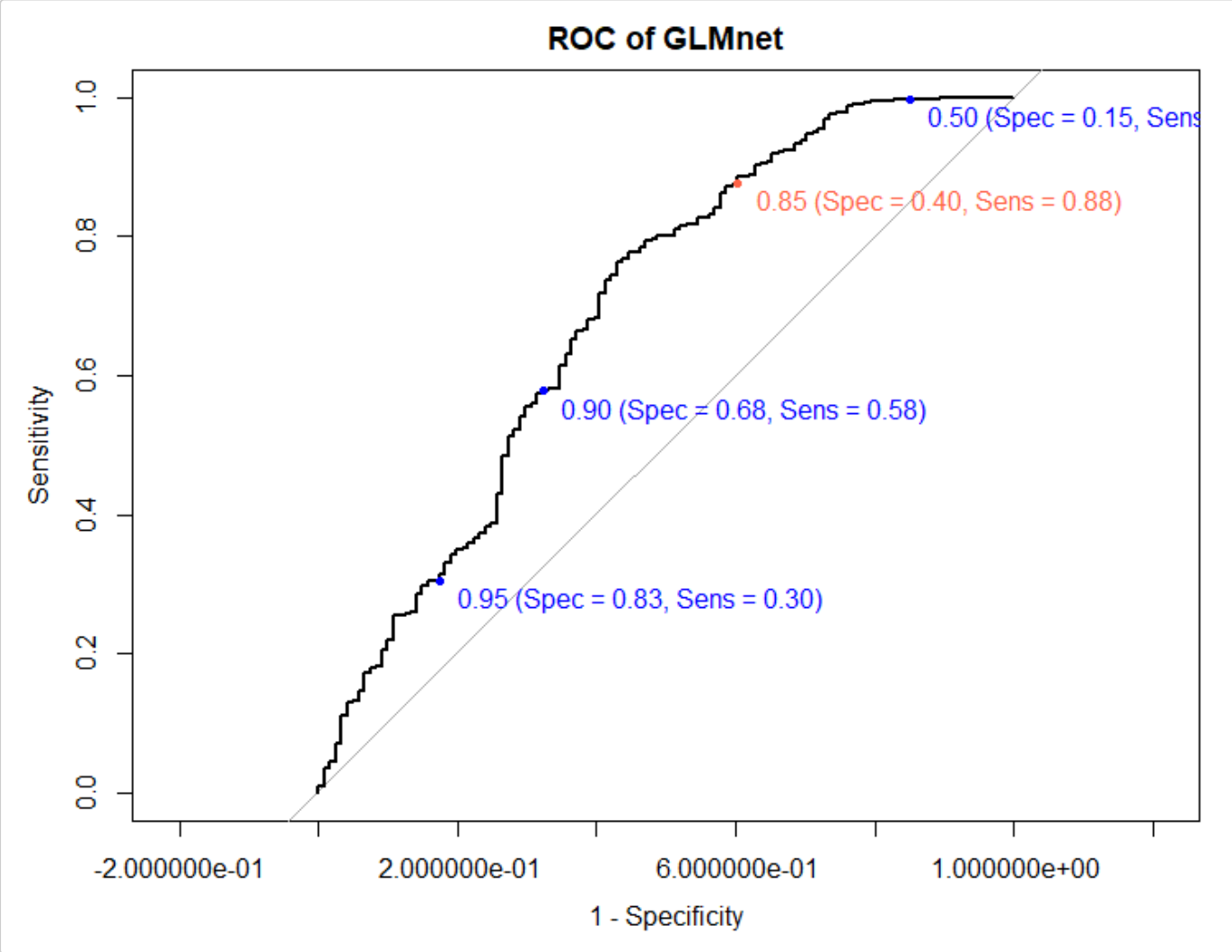


Figure 7: Possible Thresholds on GLMnet ROC Curve
*0.85 was chosen for best threshold**


ROC of GLMnet
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0
-2.000000e-01
2.000000e-01
6.000000e-01
1.000000e+00
0.50 (Spec = 0.15, Sens
0.85 (Spec = 0.40, Sens = 0.88)
0.90 (Spec = 0.68, Sens = 0.58)
0.95 (Spec = 0.83, Sens = 0.30)