US visa-class H-1B - Australia

Fabricio Martin Irabuena

6/15/2019

OVERVIEW

The US working-visa Program.

General program H-1B allows employers to temporarily employ foreign workers in the U.S. on a nonimmigrant basis in specialty occupations or as fashion models of distinguished merit and ability. A specialty occupation requires the theoretical and practical application of a body of specialized knowledge and a bachelor's degree or the equivalent in the specific specialty (e.g. sciences, medicine, health care, education, biotechnology, and business specialties, etc.).

In particular E-3 for Australia allows employers to temporarily employ foreign workers from Australia in the U.S. on a nonimmigrant basis in specialty occupations. Current laws limit the annual number of qualifying foreign workers who may be issued an E-3 visa to 10,500 Australian nationals seeking temporary work in specialty occupations.

In this analysis we focus on the Australian case. Where, we uses different *models* to predict the application's results given for each application in 2018, that is called *CASE STATUS* and we will be using *Accuracy* as a performance messure.

The base *data set* included 52 variables with above 12 thousands obserbations, some of the variables without a clear predictive power, like the ones related with post-codes and phone-numbers will be removed.

For more details on what each field means follow this link: https://www.foreignlaborcert.doleta.gov/pdf/PerformanceData/2018/H-1B_FY18_Record_Layout.pdf

The Method used for the analysis follows the steps:

1 Preparing the data.

2 Data exploration and visualization.

3 Presenting the models and evaluating the results.

4 Cross validation.

5 Final evaluation of the model's predictions on the *test set*.

PREPARING DATA

The original *data set* was constructed by selecting *E-3 Australian* cases, Where each observation reflects the details in any sigle request.

data_aus

```
## [1] 12566    39
```

We divide the *data aus* in a *train set* set with the 80% (the all the known data) and a *test set* with 20% of data (it will be consider as unknown data and will only be used for the final evaluation).

A new fiel is added *contract_days* = *EMPLOYMENT END DATE - EMPLOYMENT START DATE*

```
## [1] "dimentions"
```

train_set

```
## [1] 10052    40
```

test_set

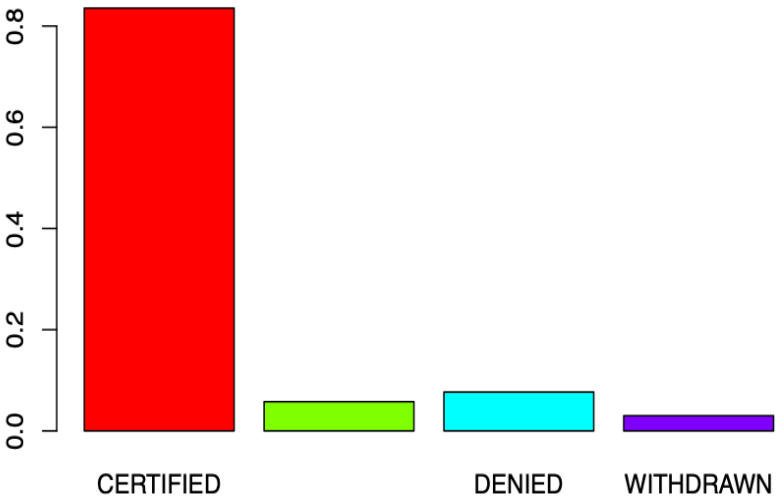
```
## [1] 2514   40
```

DATA EXPLORATION AND VISUALIZATION

Here are shown the *variable names*, how they are compounded and their *class*

```
##                      var_name uniques-factors     class
## 1                 CASE_STATUS               4 character
## 2              CASE_SUBMITTED             624   numeric
## 3               DECISION_DATE             273   numeric
## 4         EMPLOYMENT_START_DATE             699   numeric
## 5           EMPLOYMENT_END_DATE             814   numeric
## 6                EMPLOYER_STATE              56 character
## 7             EMPLOYER_COUNTRY               3 character
## 8            EMPLOYER_PROVINCE              55 character
## 9  AGENT_REPRESENTING_EMPLOYER               3 character
## 10          AGENT_ATTORNEY_CITY             330 character
## 11         AGENT_ATTORNEY_STATE              48 character
## 12                    JOB_TITLE            5119 character
## 13                     SOC_CODE             319 character
## 14                     SOC_NAME             339 character
## 15                   NAICS_CODE             898 character
## 16               TOTAL_WORKERS              18   numeric
## 17              NEW_EMPLOYMENT              11   numeric
## 18        CONTINUED_EMPLOYMENT               7   numeric
## 19  CHANGE_PREVIOUS_EMPLOYMENT              15   numeric
## 20           NEW_CONCURRENT_EMP               3   numeric
## 21              CHANGE_EMPLOYER               8   numeric
## 22             AMENDED_PETITION               7   numeric
## 23          FULL_TIME_POSITION               2 character
## 24              PREVAILING_WAGE            4085   numeric
## 25               PW_UNIT_OF_PAY               6 character
## 26                PW_WAGE_LEVEL               5 character
## 27                    PW_SOURCE               6 character
## 28              PW_SOURCE_YEAR              10 character
## 29             PW_SOURCE_OTHER             232 character
## 30       WAGE_RATE_OF_PAY_FROM            2967   numeric
## 31         WAGE_RATE_OF_PAY_TO             748   numeric
## 32            WAGE_UNIT_OF_PAY               6 character
## 33               H1B_DEPENDENT               3 character
## 34             WILLFUL_VIOLATOR               3 character
## 35                 SUPPORT_H1B               3 character
## 36              LABOR_CON_AGREE               3 character
## 37  PUBLIC_DISCLOSURE_LOCATION               1   numeric
## 38               WORKSITE_STATE              56 character
## 39           ORIGINAL_CERT_DATE             392   numeric
## 40                contract_days             194   numeric
```

CASE_STATUS	number
CERTIFIED	8398
CERTIFIED-WITHDRAWN	580
DENIED	772
WITHDRAWN	302



PRESENTING THE MODELS

Single model approach, *Random Forest*

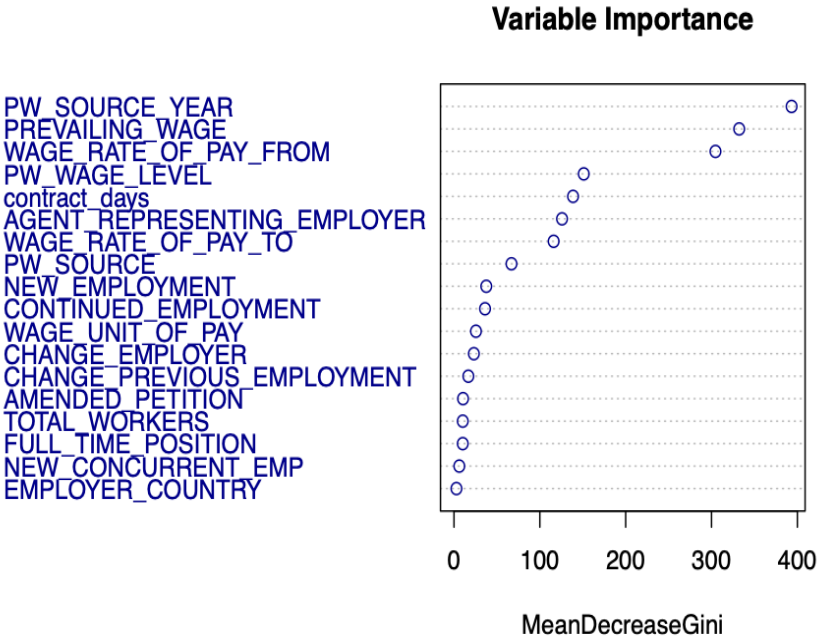
Random Forest will be used as a first aproach model, and to get a general idea of the relative impact that each of the pre-selected variable has, thanks to the *importance* function.

Note: Only the varables with less than 50 uniques-factores where selected, that is the maximun the algoritm allows. In addition, *H1B-DEPENDENT*, *WILLFUL-VIOLATOR* and *SUPPORT-H1B* where not consider due to the extended number of *NA* values within them.

Accuracy
0.9078

```
##                                 MeanDecreaseGini
## AMENDED_PETITION                       10.524946
## EMPLOYER_COUNTRY                        2.830899
## WAGE_UNIT_OF_PAY                       25.584217
## FULL_TIME_POSITION                     10.191612
## AGENT_REPRESENTING_EMPLOYER           125.845288
## PW_WAGE_LEVEL                         151.159585
## PW_SOURCE                              66.922454
## PW_SOURCE_YEAR                        393.266983
## NEW_CONCURRENT_EMP                      6.173205
## CHANGE_EMPLOYER                        23.137900
```

```
## CHANGE_PREVIOUS_EMPLOYMENT          16.696583
## NEW_EMPLOYMENT                      37.610272
## CONTINUED_EMPLOYMENT                36.164348
## PREVAILING_WAGE                    332.216098
## TOTAL_WORKERS                       10.198364
## WAGE_RATE_OF_PAY_FROM              304.594663
## WAGE_RATE_OF_PAY_TO                116.071373
## contract_days                      138.665254
```



Now we are able to removed the variables with a very low impact on the model. In this scenario, we will pick out those with a *MeanDecreaseGini* below 10.5

After removing them, we can see that even with less variables the *Accuracy* of the model remains at the same level or even increases. In other words, a simpler model with the same predicting power.

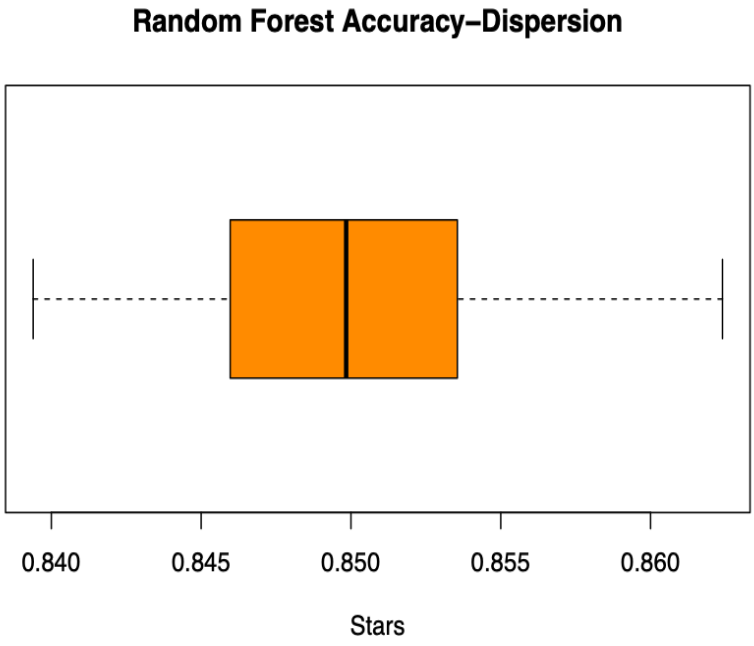
Accuracy
0.9129

Now, to get an better idea of the model's behavior, we take a look at the crossed table between the model's prediction and the actual data in the *train set*.

```
##                       CASE_STATUS
## pred_rf_opt            CERTIFIED CERTIFIED-WITHDRAWN DENIED WITHDRAWN
##   CERTIFIED                 8371                 255    328       240
##   CERTIFIED-WITHDRAWN          9                 324      4        15
##   DENIED                      16                   1    422         5
##   WITHDRAWN                    1                   0      0        41
```

Croos Validation* on *Random Forest

We are taking many new *samples* of the *data set* to evaluate the model and find out what would be the expected *Accuracy* on each new sampled *test set*, the selected *size* is the same as in original setup. Within this case 50 new *samples* were taken.



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.8394	0.846	0.8498	0.8496	0.8535	0.8624

Multiple models approach, *Caret Package*

Thanks to the *Caret Package* features, we are able to train different models at the same time. The ones used here are:

```
## [1] "Support_Vector_Machine_Radial"   "ranger"
## [3] "Weighted_Subspace_Random_Forest" "k-Nearest Neighbors"
```

```
## [1] 4
```

svmRadial	ranger	wsrf	kknn
0.8666	0.856	0.9141	0.9022

Croos Validation* on *Caret approach

We are taking many new *samples* or *test set* from the *train data* to evaluate the all the models and find out what would be the expected *Accuracy* on each new sampled *test set*, again the selected *size* resembles the

original setup. Within this case 50 new *samples* were taken.

	svmRadial	ranger	wsrf	kknn
Mean-Accuracy	0.85	0.8379	0.8949	0.8821

RESULTS

Finally, we are able to evaluate our best-performance's model *wsrf* using the original *test set*. Which, remains yet as unknown data, and reveals the results under a possible real scenario. *0.8854*

CONCLUSION

After following the steps in the Analisys, we have improved the base line of a *Single Model* with a *Random Forest* for predicting *CASE STATUS*, going through a *Multiple-models* approach, thanks to the *Caret Package* and finally we concluded that the highest *Accuracy* was reached by the *wsrf* model, with above 88%.

Please follow this link to have access to the project files: https://github.com/AmadoLabX/Data_Science_HX