Capstone Project 1- H1B ML

Sheila Torrico - October 2018 Cohort

Problem Statement

It is known that many US open job positions can not be filled due to the lack of qualified resources within the United States. One option that companies are using to solve this shortage is by sponsoring foreign qualified professionals via H1B visas. Companies will submit applications on behalf of these foreign workers.

The main outcome that companies and individuals are hoping for is Certified (approved). Besides it, the other potential outcomes are Rejected/Denied/Withdrawn/Pending.

Project Goal

The goal of my project is to answer the following question: **Is it possible to predict if an H1B application will be Certified or not given the selected company sponsor/job title?**

The answer to this question will help individuals use this knowledge to assess which company/position they should target so they can make early plans for example around their job strategy or which study field to pursue. It could also help companies that are building a mobile app, which will make this prediction, to determine which questions (variables) should be included in the interface. For this project, I am assuming that I am helping the business team with this decision.

Dataset

The historical data that I will use to answer this question will cover at last fiscal year (Oct 1, 2017- Sept 30, 2018) of H1B visas applications and is available for collection at the Department of Labor.

Solution Approach

1. Phase 1: Data Wrangling of dataset
2. Phase 2: Perform EDA on the collection data to identify most likely predictor from potential inputs. Conduct Inferential Statistical Analysis on representative sample of dataset with focus on:
    a. Sponsor Company
    b. Job Title Position
    c. Potential Salary Range
    d. Duration
3. Phase 3: Select focus features (predictors) and perform feature engineering
4. Phase 4: Select ML algorithms to predict if application will be certified
5. Phase 5: Review results and present insights
6. Phase 6: Call to actions

Deliverables

1. Code Jupyter Notebooks (Cleaning/EDA/ML algorithm)
2. Report summary (word/presentation deck)

Phase 1

The first phase on my project consisted of several steps such as collecting the data, reviewing it and deciding how to transform it for the purpose of building a model that will be able to predict if a visa application will be Certified or not.

Dataset

From the Department of Labor, I collected the H1B-visa-applications file from the past fiscal year (Oct 1, 2017- Sept 30, 2018). The file was available in .xlsx format.

Data Review

Before uploading the file to a Jupyter notebook, I did a preliminary analysis in Excel by using a sample of 2000 observations. The review goals were to understand the type of information that each of the 52 variables had, the quality of the observations, and if there were any mandatory vs optional variables.

The results gave me an insight on the data types I should expect to see once the file is available in a Jupyter notebook and a glimpse of the variables that might have missing observations or outliers. For the next step, I uploaded the sample file in Jupyter notebook and used available functions from pandas to confirm the data types and that all variables and corresponding observations were successfully uploaded.

Given the positive outcome, I uploaded the complete file (dataset size ~600k) and sure enough all my variables and corresponding observations were in the notebook. The variables were divided into the following data types: integer, float, datetime and object.

Data Wrangling

To continue with the transformation process of the data, I divided my variables into non-object and object data types so I could apply the appropriate techniques.

For the non-object variables, I identified the variables that had any missing observations and calculated the percentage of missing observations. Also, I used statistical functions that measured central tendencies. Through them, I was able to identify variables that might have outliers. For variables with outliers, I performed an additional analysis to find out potentials reasons/patterns. I.e., in the case of the wage variable, the large standard deviation was the result of having such disperse values that ranged from 0 to 1,000,000,000.

I created a separated dataframe for variables with an object data type so I could assign them to the categorical data type. Before performing such step, I identified all variables that had any missing observations (NaN) and calculated the corresponding percentage. Then, I extracted a subset from these variables to perform further analysis and identify what potential values could be used to fill the NaN observations.

Results

From performing these steps, the dataset was enhanced by:

- Removing variables that had at least 90% of missing observations.
- Filling the NaN values of variables that had at least 95% of all observations.
- Not filling or removing variables that have 30% to 60% of missing observations. Instead, will look for a ML algorithm that could handle large number of NaN values.
- Transforming categorical variables to new numerical variables.

Phase 2

For the second phase, it was about performing Exploratory Data Analysis on the transformed dataset to get some initial findings on my focus variables like Duration (processing times), Job Title, Employers and discrete variables such as if employer was a willful violator or the position was full time or not. Note, Duration variable was calculated by taking the difference of decision date and case submitted date.

As part of the EDA, I also performed an statistical analysis to determine if the Duration variable followed a Normal Distribution. Used a RSS sample size of 7k observations from the original dataset.

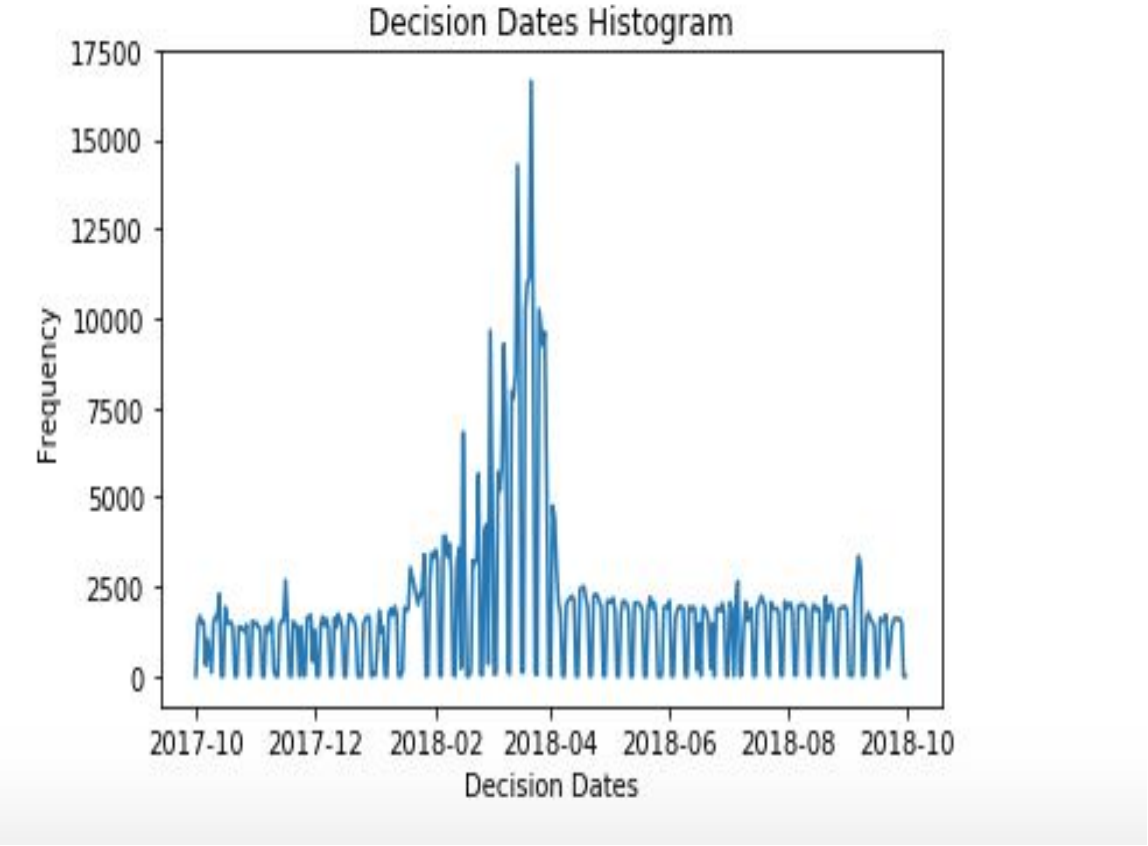
EDA Findings

1. A rough estimate shows that at least 80% of the submitted applications obtained the Certified status

```
CERTIFIED               579449
CERTIFIED-WITHDRAWN      45004
WITHDRAWN                21280
DENIED                    8627
Name: CASE_STATUS, dtype: int64
```

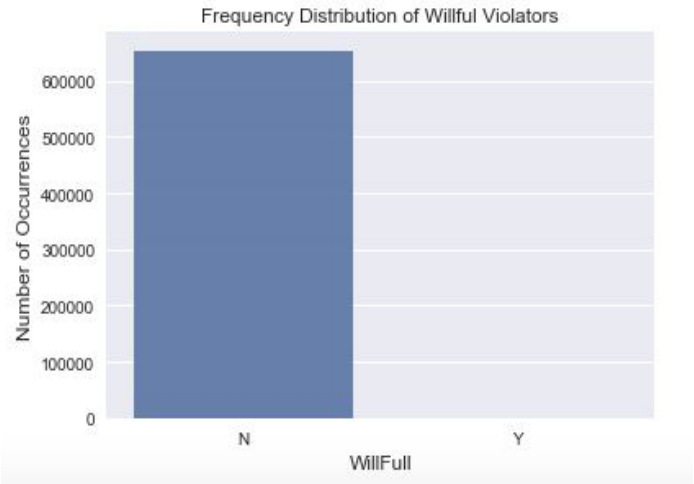
2. It can be observed from Decision Dates Histogram the existence of a peak during February and April. This peak indicates that a large number of applicants received an answer on the status, Certified or not, of their applications.

Fig.1- Decision Dates Histogram



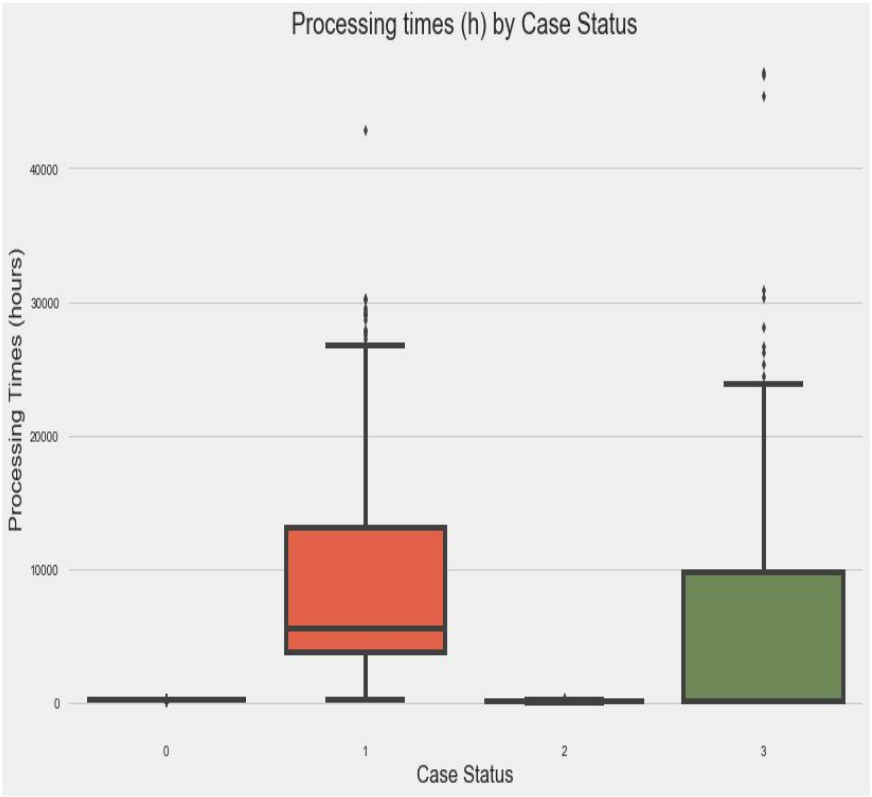
3. It is confirmed that majority of sponsors (employers) are compliant with the H1B visa requirements by observing Fig. 2.

Fig. 2. Willful Violators.



4. An interesting finding is that for applications that ended up being Withdrawn(Cat 3), there are many processing times along the whisker that are outside of the highest observation as observed in Fig 3. The second one finding, is that highest observation of the set belongs to applications that ended up being Certified-Withdrawn (Cat 1). Clear Processing times for applications that are either Certified (Cat 0) or Denied (Cat 2).

Fig. 3. Processing times by Case Status



Inferential Statistics Results

5. I was able to conclude that observations belonging to the Duration variable (processing times) tend to a Normal Distribution. As next step, performed a Z test hypothesis testing on the Duration variable to answer the question if it is possible that the processing time can be less than Average of 42.65 days.

   Ho: μ <= 42.65 days
   Ha: μ >= 42.5 days

   The result indicated that I do not reject the Null Hypothesis with a confidence of 95%. This result opens the door to potential improvements on the processing time that could help improve applicant experience.

6. An interesting finding is that the job category (55) equal to Software Developers, Applications (based on SOC code) is the job that most employers are trying to fill. Thus, takes has the longest processing time.

7. Another interesting find is the 2 categories (258 and 259) at the bottom of the list correspond to Airline Pilots, Copilots, and Flight Engineers and Bus Drivers, Transit and Intercity respectively with a equal processing time of 1 hour.

8. The resulting correlation on the variables: job titles (counts) and how long it takes to process them is positive and very close to 1. Thus, both variables have a stronger linear relationship.

Phase 3

Feature engineering

Given the results from the EDA phase, it was time to build a predictive model that could answer the question that an H1B application could be Certified or not using some of my focus variables, job title, employer as predictors from the enhanced dataset. Based on EDA results, processing time (duration) was also added as a feature (predictor).

From additional research on the application process, it became clear that wage and wage level needed to be included as features since the U.S. Department of Labor generally requires that hiring of a foreign worker will not adversely affect the wages and working conditions of U.S. workers in similar jobs. The information on wages would not be meaningful without knowing the type of position (full time vs not full time) and the area of intended employment, represented by the worksite postal code feature. Thus, 7 features were selected to predict the label (Case Status).

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 654359 entries, 0 to 654359
Data columns (total 7 columns):
EMPLOYER_NAME            654359 non-null object
JOB_TITLE                654359 non-null object
FULL_TIME_POSITION       654359 non-null object
DURATION                 654359 non-null float64
WAGE_RATE_OF_PAY_FROM    654359 non-null float64
WORKSITE_POSTAL_CODE     654359 non-null object
PW_WAGE_LEVEL            616141 non-null object
dtypes: float64(2), object(5)
memory usage: 39.9+ MB
```

Before going ahead and start using appropriate Machine Learning models, it became evident that I needed to simplify 3 categorical predictors. Why? Employer variable has over 200 unique employers that submitted applications, which translates into having 200 potential categories or classes within this variable!! Something similar is observed for the Worksite postal code feature. For the purpose of simplification, decided to focus on the top 50 employers, worksite postal codes and job titles. Just then, I applied one-hot encoding to all required categorical predictors. At the end, the features dataset went from 7 to 161 variables.

Phase 4
Machine Learning Models

Considering that the outcome of my label variable is a binary, the following classification models were selected: Logistic Regression, Decision Tree, Random Forest and Gradient Boosting. The features and label data were splitted into a training set, test set and validation set.

The training data was used to build the model. Ie. Random Forest. Then, the model was used on the test data to evaluate its performance and finally, the validation data was used for hyperparameter tuning.

Phase 5
Machine Learning Results

The below summary table compares the accuracy for the different classifiers and most importantly, the AUC score. The AUC score indicates the effectiveness of the classifier used. Note that the accuracy scores from the models are within 89 - 98% range and the AUC scores are almost perfect for the Random Forest and Gradient Boosting classifier

Model	Accuracy	AUC Score
Logistic Regression	89%	78%
Decision Tree	98%	95%
Random Forest	99%	96%
Gradient Boosting	99%	96%

The top features that contributed to the Random Forest and Gradient Boosting classifiers were determined and not surprisingly, Duration feature comes as the most import one.

Analysis

The results seem too good to be truth, particularly the AUC scores and further research confirmed that when such scores are unrealistically high, is a clear indication that my model is suffering from Target Leakage. Realizing that when creating the Duration feature, I included information that was not going to be available when making the prediction: Decision Date. Basically, the Department of labor filled the Decision Date for the application once the decision was made.

Another potential explanation that my results seemed unrealistic, is that my label variable is a clear example of imbalanced data. There ratio of applications that were certified to applications that were not was 4/1 so the classifiers were more biased toward the majority class or “Certified” status.

Contingency Plan

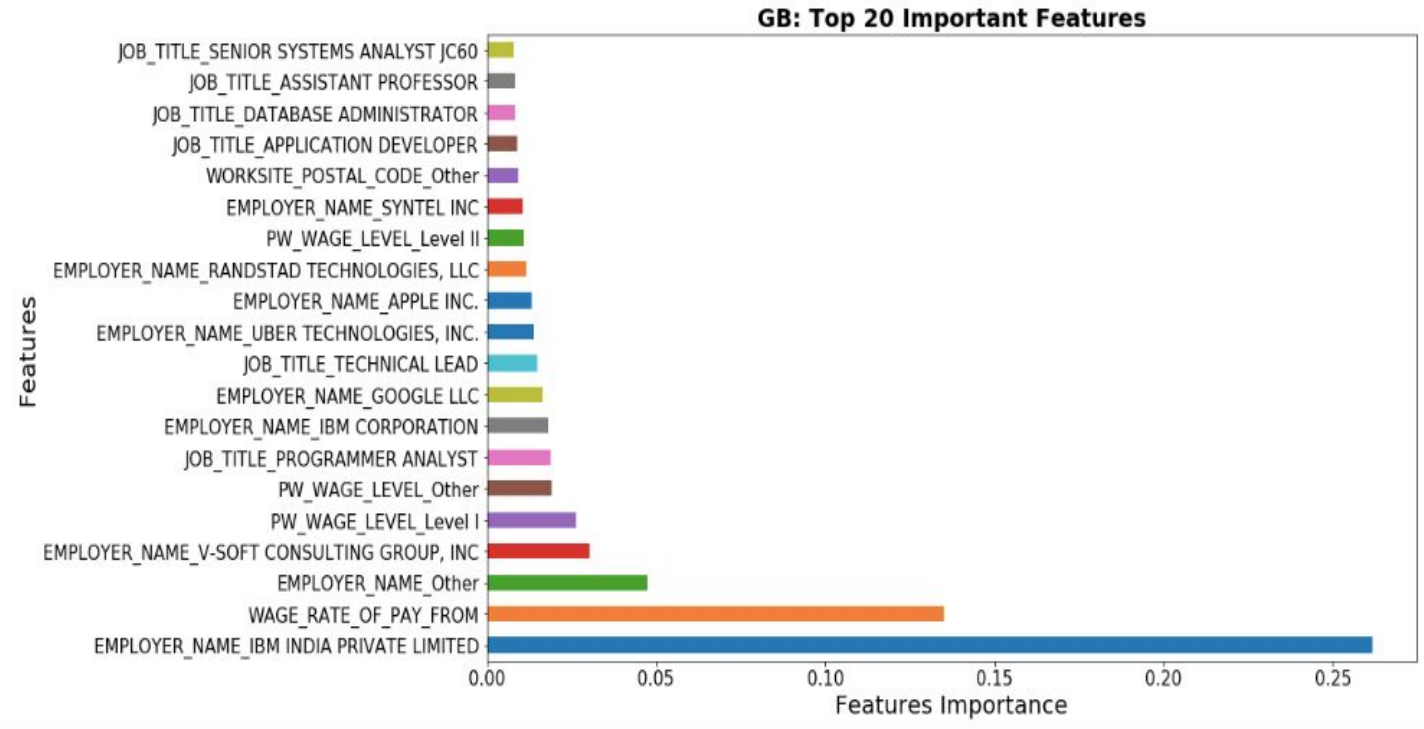
To fix the Target leakage issue, I excluded Duration from the features dataset. Performed another ML iteration using the improved features dataset and corresponding label variable. Applied the same classifiers and calculated their AUC scores.

The below summary table shows that now the scores are more realistic and at least they are not completely ineffective. Among all classifiers, Gradient Boosting was the most effective with a score of 67%.

Model	AUC Score
Logistic Regression	56%
Decision Tree	64%
Random Forest	66%
Gradient Boosting	67%

The top 20 features that Gradient Boosting used when modeling the data are seen below in Fig 3.

Fig 3: Top 20 Features identified by Gradient Boosting



Interestingly enough, one of the top features belonged to an underdog Employer: V-Soft Consulting Group, Inc. The next top features were: the level type I for the wages, the Programmer Analyst job title and the wages for the different jobs.

Phase 6

Conclusions

Some key takeaways from applying the models to try to answer if it is possible to predict if an H1B application will be Certified or not are:

- Employer is a valid predictor variable and selecting the appropriate employer could help the likelihood of getting the Certification status. Just 2 employers from the top 3 employers with highest number of application submitted made it as top contributor for the model.
- Job title is also a valid predictor variable. The top job title that these employers are trying to fill is for Programmer Analyst which is typically entry level and lower-paid. Although this job is in the top 3 job titles, it does not guarantee that an application will be certified.
- The wage that an employee will be paid and the wage levels are valid predictors. Wage Level in particular helps determine if the job is considered a speciality one.

Call to Action Business Group:

1. Model showed during first iteration that duration variable was creating noise. [Action] Agree on definition of Duration.
2. Besides the 3 features business wanted to evaluate, model identified Worksite zip code and Wage level as determining predictive features. [Action] Make a decision if these 2 variables should be part of Product.
3. Results show that a potential spin-off of Product could be to display “underdog”employers to users. [Action] Research the viability of this model suggestion based on results

Call to Action Data Science Group

1. To improve model effectiveness, the options are to balance the data or include another feature that already exists in the dataset or construct another one.
    a. Potential Features: Prevailing wage variable or baseline wage or construct the Employment Duration by using the employment start date and end date
2. Before deploying the model, agreement is needed on how to deal when the model indicates that an application will be certified when is not (what is the cost of a false positive?).
    a. Should this be penalized? [Action] Obtain agreement on the type of penalization to be applied.