CAPSTONE PROJECT 1: PREDICT IF H1B APPLICATION WILL BE CERTIFIED OR NOT

SHEILA TORRICO

Background

- Many US open job positions can not be filled due to the lack of qualified resources within the United States.
- One alternative for companies is sponsoring foreign qualified professionals via H1B visas.
- The main outcome that companies and individuals are hoping for is Certified (approved). The other potential outcomes are Rejected/Denied/Withdrawn/Pending.
- Potential applicants would like to get some insights on some employers/job title combinations to target so their could obtain H1B visas.

Project Goal

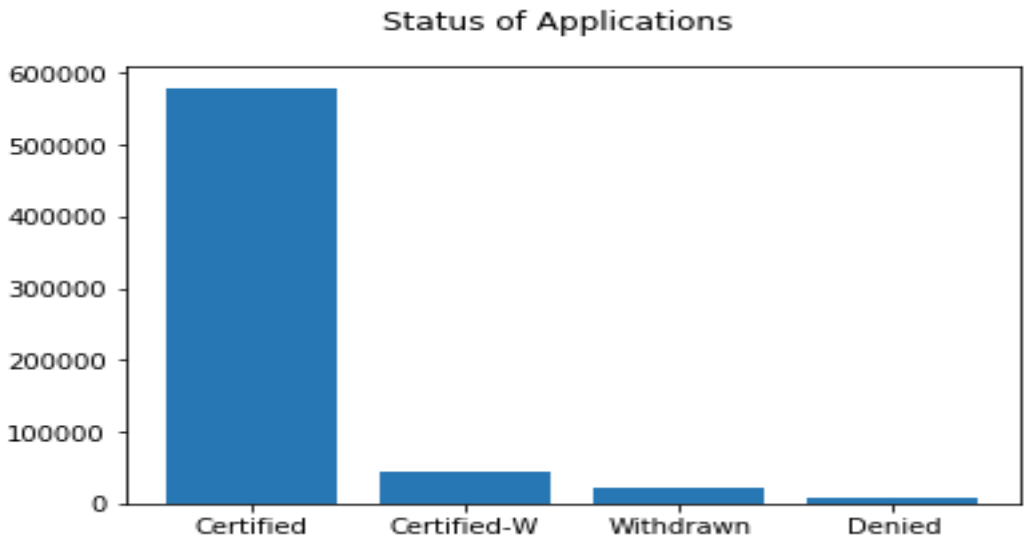
- Business team goal is to launch an app that would predict if an application will be Certified or not. They would like to know what information (variables), should be asked to potential users.
  - The goal of the project is to answer the following question: Is it possible to predict if H1B application will be Certified or not given some set of variables (features) such as company sponsor/job title/ wage/duration?

Solution Approach

- Phase 1: Perform Exploratory Data Analysis
- Phase 2: Evaluate Results
- Phase 3: Report on insights (pattern/trends) and Call to action

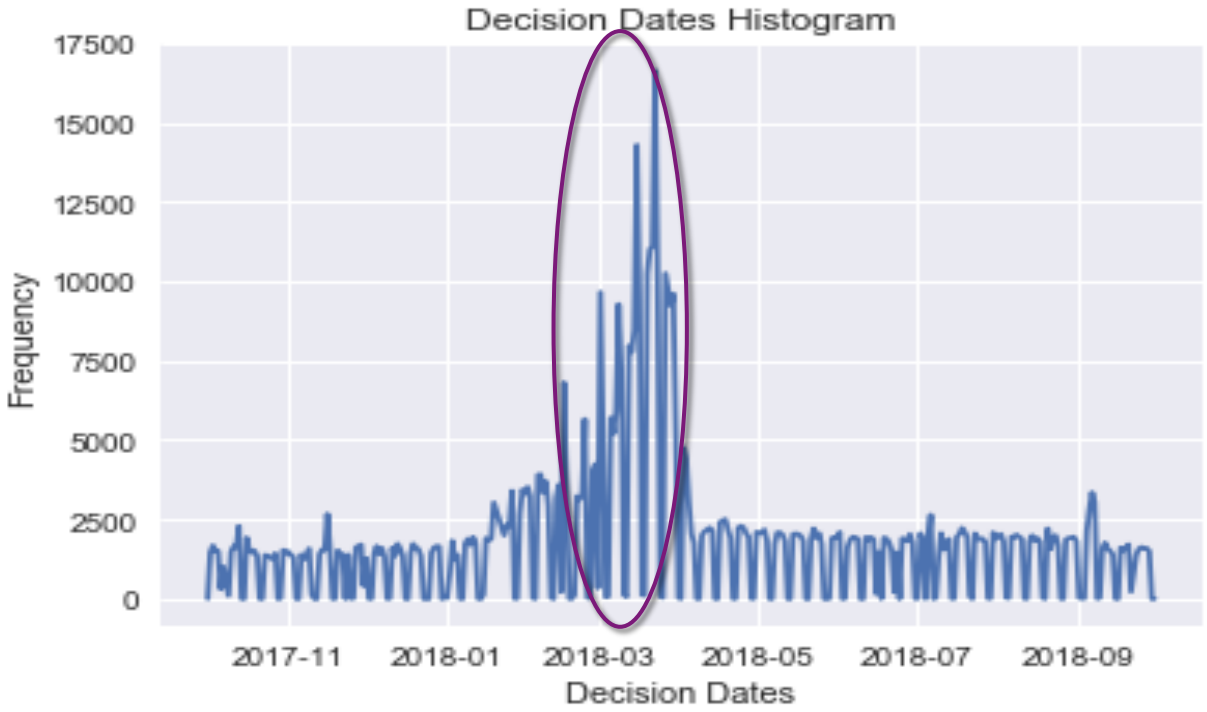
Phase 1: Exploratory Data Analysis

- Dataset used covers last fiscal year (Oct 1, 2017- Sept 30, 2018) of H1B visas applications. Available for collection at the Department of Labor.
- Quick glimpse of outcomes from submitted applications: at least 80% obtained Certified status



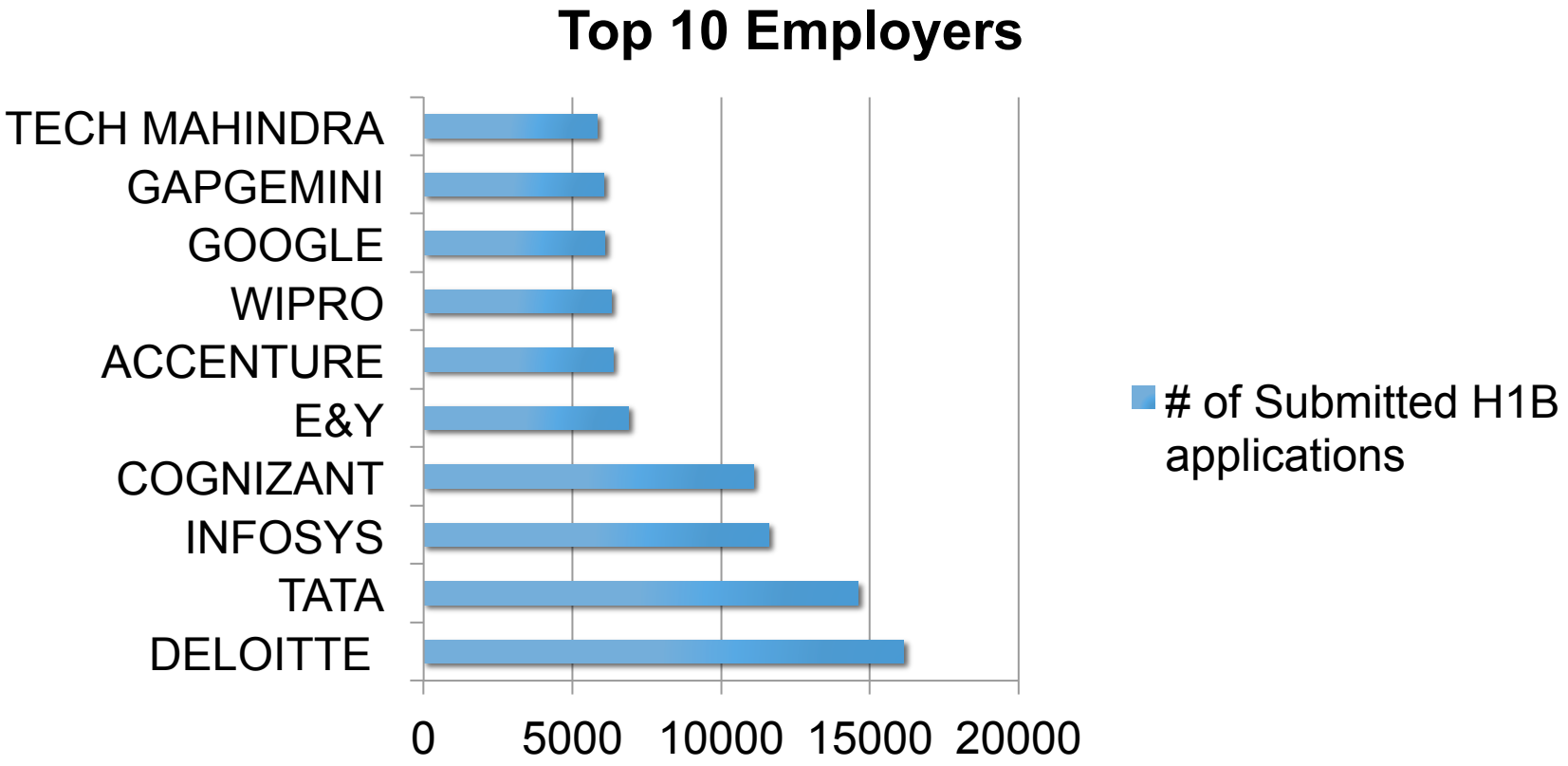
Phase 1: Exploratory Data Analysis

- A large number of applicants received an answer on the status of their applications during February-April. Expect more user traffic to check if our product correctly predicted the status



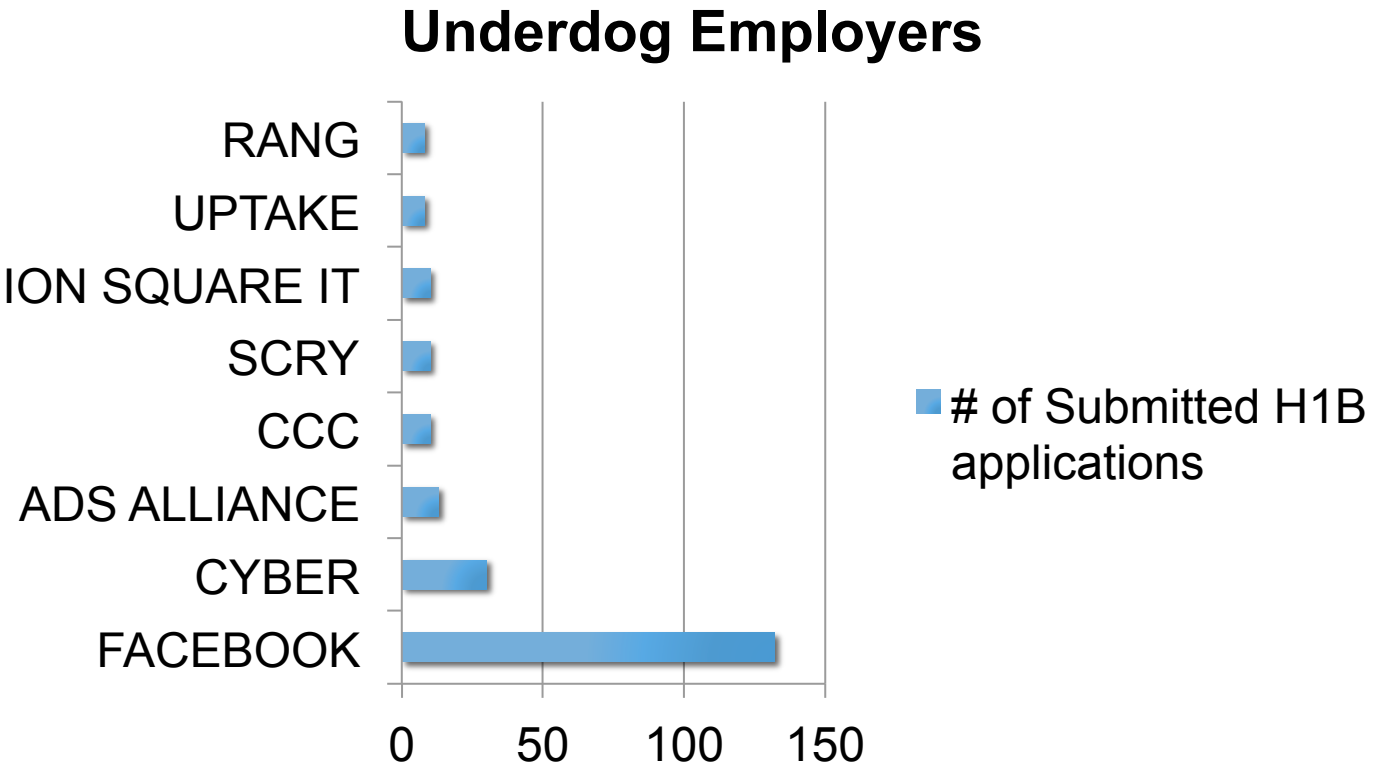
Phase 1: Exploratory Data Analysis

- Employers with the highest number of submitted applications



Phase 1: Exploratory Data Analysis

- Besides known employers such as Facebook, what other employers an applicant can target if he/she is a Data Scientist.



Phase 2: Evaluate Results

- Built 4 different models that could predict if an application will be certified or not.
- Gradient Boosting model obtained the best AUC Score.

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

Phase 2: Evaluate Results

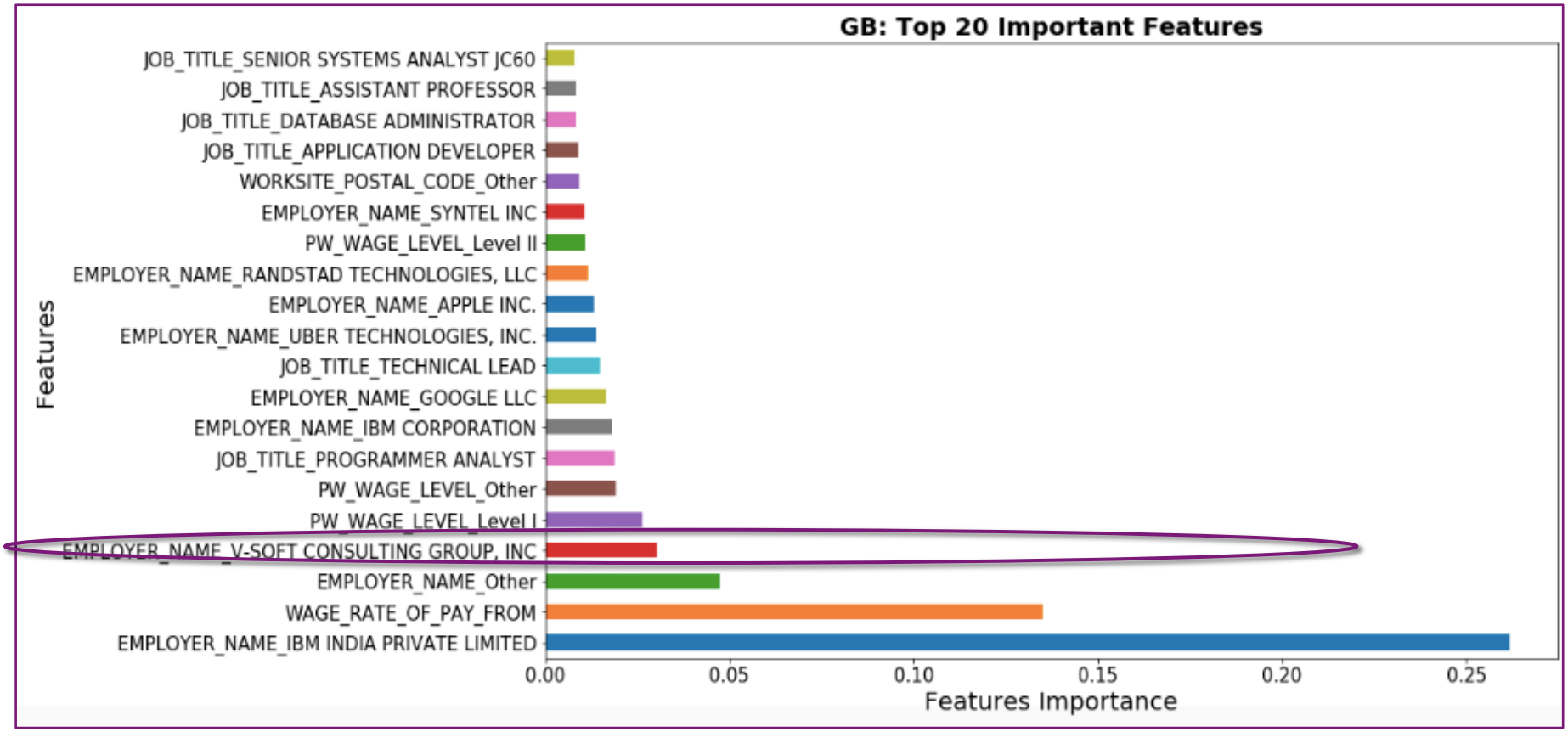
- Classification Report for Gradient Boosting Model

Status	Precision	Recall	F1-Score
Not Certified	.65	.04	.08
Certified	.89	1.00	.94
Avg/total	.86	.89	.84

- For Certified applications:
  - Gradient Boosting Model correctly predicted 89% of the time that an application will be certified.
  - Recall score of 1 was obtained, which means the model correctly classified all available applications.

Phase 3: Insights and Call to action

- Top features that Gradient Boosting used when deciding if an application will be certified or not.
- Underdog employer: V-Soft Consulting group made it to top 5.



Phase 3: Insights and Call to action

- Model identified 5 variables as key predictors: Employer, Job Title, Wage Rate of Pay from, Wage Level and Worksite Postal Code.
- Not a surprise that Employer was a key predictor variable, but surprisingly 2 **underdog** employers were considered: V-Soft Consulting & Syntel, Inc.
- The proposed wage that an employer will pay for a job title is key.
- Wage Level helps determine if the job is considered a speciality
- Although Programmer Analyst made it to the top 10, is not guarantee the application will be certified.

Phase 3: Insights and Call to action

- Call to action to Data Science group:
  - 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?). Should this be penalized? **[Action]** Obtain agreement on the type of penalization to be applied.

Phase 3: Insights and Call to action

- Call to action to Business/Product group
  - Model showed during first iteration that duration variable was creating noise. **[Action]** Agree on definition of **Duration**.
  - 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.
  - 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.