Predicting the Outcome of H-1B Visa Applications

Beliz Gunel, Onur Cezmi Mutlu
{bgunel, cezmi}@stanford.edu

Problem

- H-1B is a type of non-immigrant visa for foreign nationals with specialized knowledge.
- US Immigration Service grants 85,000 H-1B visas every year, even though the number of applicants far exceed that number. The selection process is based on a lottery, hence how the attributes of the applicants affect the final outcome is unclear.
- We aim to build a prediction algorithm which could be a useful resource both for the future H-1B visa applicants and the employers who are considering to sponsor them.

Dataset & Features

- H-1B Visa Petitions 2011-2016 dataset' from Kaggle was used. It included around 3 million datapoints with 9 initial features.

CASE_STATUS	EMPLOYER_NAME	SOC_NAME	JOB_TITLE	FULL_TIME_POSITION	PREVAILING_WAGE	YEAR	WORKSITE
DENIED	ARAPURA INC	GENERAL AND OPERATIONS MANAGERS	OPERATIONS MANAGER	N	60000.0	2016.0	SOUTHPORT, CONNECTICUT
CERTIFIED	CITIUSTECH INC	COMPUTER SYSTEMS ANALYSTS	COMPUTER PROGRAMMER ANALYST	Y	89086.0	2016.0	ORANGE, CALIFORNIA

- After pre-processing, the final features were: {APPS_PER_COMPANY, COMPANY_SUCCESS_RATE, APPS_PER_SOC, SOC_SUCCESS_RATE, YEAR, FULL_TIME_POSITION, PREVAILING_WAGE, (onehot-k), WORKSITE (onehot-k)}
- Final number of features was 64.
- We used the «CASE_STATUS» as our labels.

Results & Discussion

- We split the training, dev and test sets 60:20:20 , and tuned the hyperparemeters. Then, we retrained the best-performing models on 80% of the data and tested on the remaining 20%.
- Due to the inherent bias in dataset towards the «APPROVED» label, we also tested the accuracy on balanced test data (data set with rougly equal number of each label).

	Train acc. (balanced)	Test acc. (balanced)	Train acc. (unbalanced)	Test acc. (unbalanced)
Naive Bayes	94%	72%	94%	94%
Log. reg. w/ l1	98%	74%	98%	98%
Linear SVM w/ ElasticNet	98%	78%	98%	97%
MLP	99%	70%	99%	98%
Neural Net.	98%	76%	98%	96%
Neural Net. w/ l2	98%	82%	98%	97%

- In general, there is a big variance between training and test accuracy. However, the results show that including l1 regularization in the model increases the test accuracy. This is due to the irrelevancy of some features to the output, considering l1 regularization enforces sparsity in the feature vector.
- Neural Network model with l2 regulatization outperformed all the other models.

Models

- Several different models were trained using both stochastic and minibatch gradient descent methods:
  - Logistic regression
  - SVM
  - Naive Bayes
  - Multilayer Perceptron
  - Custom Neural Network
- For all models, the following methods were utilized to tune the hyperparameters:
  - Polynomial and Gaussian kernels
  - Regularization
    - l_1, l_2 and ElasticNet

Future Work

- Converting more features such as SOC_NAME into one-hot-k representation could help achieve better accuracy.
- Depth of the neural network, number of neurons at each layer and the network architecture could be adjusted.
- Instead of using the given EMPLOYER_NAME and SOC_NAME features directly, more informative features such as Standard Industrial Classification codes of the companies could be created through web crawling.

References

1. H-1B Visa Petitions 2011-2016 | Kaggle. [Online]. Available: https://www.kaggle.com/nsharan/h-1b-visa/data. [Accessed: 20-Oct-2017].
2. "High-skilled visa applications hit record high" CNNMoney. [Online]. Available: http://money.cnn.com/2016/04/12/technology/h1b-cap-visa-fy-2017/index.html. [Accessed: 20-Oct-2017].
3. H-1B Fiscal Year (FY) 2018 Cap Season," USCIS. [Online]. Available: https://www.uscis.gov/working-united-states/temporary-workers/h-1b-specialty-occupations-and-fashion-models/h-1b-fiscal-year-fy-2018-cap-season.[Accessed: 20-Oct-2017].
4. " Predicting Case Status of H-1B Visa Petitions." [Online]. Available: https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a054.pdf.
5. " H-1B Visa Data Analysis and Prediction by using K-means Clustering and Decision Tree Algorithms." [Online]. Available: https://github.com/Jinglin-LI/H1B-Visa-Prediction-by-Machine-Learning-Algorithm/blob/master/H1B\%20Prediction\%20Research\%20Report.pdf.