School of Business
D'Amore-McKim
Northeastern University

The value of international students to the United States. Probability of getting a non-immigrant visa.

Project Report
MISM6212 Data Mining and Machine Learning
MS in Business Analytics - Class of 2021
D'Amore-McKim School of Business
Northeastern University

Team 11
Zinaida Dvoskina | Kirill Ilin
Johnathan Conley | Cindy Ye Fung

Part 1 - Business Question:

Northeastern University is among the top US colleges for international students under the F-1 visa. Our MSBA cohort is not an exception - it represents 16 foreign countries, so many of us are directly affected by the changes in immigration policies and are very vulnerable to them. As we're towards the end of completing our degree, many will be seeking career opportunities in the United States; however, this can be challenging due to several macro and micro factors.

International students are beneficial to the US economy and create thousands of workplaces. However, the reality is that the job market is challenging as "international students need approval from the U.S. government before they can work [...]. That's a disincentive for some employers" (Chen, 2019). A common work authorization program for international students is the Optional Practical Training (OPT) program directly related to the student's area of study. Eligible students can receive up to 12 months of OPT employment authorization, and students in the STEM field can apply for a 24 months extension of the OPT period (USCIS, 2020). The next step after OPT is for international students to obtain an H-1B visa which allows a foreign national to work temporarily in the U.S. for up to six years; however, it must be petitioned and paid by the employer (USCIS, 2020).

Recently the immigration issues - including non-immigrant visas - have become very political, so policy creation does not consider the real impact. Within our project, we'd like to analyze publicly available data on the U.S. non-immigrant visa acquisition to emphasize it and explain why the number of visas should not be declining like it is.

- **Question:**

Seeking to understand if various factors such as the passport that the student holds, local location of the applicant, type of visa the student is seeking, the political party in office, wage rate, job title, etc., determines the likelihood of an international student obtaining a working visa.

Part 2 - Data Sources:

To conduct our research, we are using publicly available data mainly from governmental websites with qualitative data that includes immigration law or student majors data and quantitative data for wages and dates.

The first source is from the U.S. Citizenship and Immigration Services USCIS, which has data on the number of visas issued per country, category, the political party in office, and year, seen in Exhibit 1 below:

Column Name	Column Description
Fiscal Year	Includes the years between 1997 to 2020
Region	Locations for Africa, Asia, Europe, North America, Oceania, Other, South America
Country	Includes 208 different countries
Visa Category	Includes 91 different visa types
Number of Visas	Number of visas approved
Party in Office	Democratic or Republic party during visa approval

The second source is from the U.S. Department of Labor has Performance Data that includes information about employment-based immigration applications from the Office of Foreign Labor Certification's case management systems, seen in Exhibit 2 below. The dataset is cumulative for the fiscal year and contains columns showing records provided by employers such as an applicant's received dates, decision dates, the most recent date a case determination decision was issued, etc.

Exhibit 2:

Column	Column Name	Description
A	Year	Fiscal Year
B	Status	Status associated with the last significant event or decision. Valid values include "Certified", Denied".
C	LCA_CASE_SUBMIT	Date and time the application was submitted
D	DECISION_DATE	Date on which the last significant event or determination was issued by OFLC.
E	LCA_CASE_EMPLOYMENT_START_DATE	Beginning date of employment
F	LCA_CASE_EMPLOYMENT_END_D	Ending date of employment

	ATE	
G	LCA_CASE_EMPLOYER_NAME	Employer's name
H	LCA_CASE_SOC_CODE	The Standard Occupational Classification (SOC) code which classifies workers by occupational groups
I	LCA_CASE_SOC_NAME	Title of the SOC occupational group
J	LCA_CASE_JOB_TITLE	Job title
K	LCA_CASE_WAGE_RATE_FROM	Employer's proposed wage rate
L	WORK_LOCATION_CITY1	Address information of the intended are in which the foreign worker is expected to be employed (city of the job opening)
M	WORK_LOCATION_STATE1	Address information of the intended are in which the foreign worker is expected to be employed (state of the job opening)
N	PW_Fixed	Hourly Prevailing wage rate.

- **Data Processing:**

For the purpose of our project, we have selected specific columns from the Performance Data shown in Exhibit 2 and cleaned the data for better analysis (See Exhibit 1 for the explanation of each column). One of the significant changes is on the "Status" column, where we only include rows with "Certified" and "Denied" results. We decided to eliminate the rows with a "Withdrawn" status and changed "Certified-Withdrawn" status to "Certified" to highlight the impact of international students who completed the H1-B visa application. Furthermore, we added the "FW_Fixed'' column to standardize the prevailing wage rate to be yearly (consisting of 12 months and 51 weeks considering some holidays), and it was calculated by multiplying the rows with hourly values by 2087. The prevailing wage is the average wage paid to similarly employed workers in a specific occupation in the area of intended employment, and this column can provide valuable insights as we want to calculate the difference with the employer's proposed wage rate column in our analysis section to understand expected salary that an international student can potentially expect.

Part 3 - Analysis:

Our initial hypothesis was to determine if the party in office affects the likelihood of a foreigner obtaining a working visa. We added visualizations with the data from USCIS using Tableau, and we noticed right away that this variable alone is not a good predictor whether a student obtains a work visa or not. This is visualized on Exhibit 3 seen below, where there is not a strong trend to justify this hypothesis. Nevertheless, the data is useful to provide insights about the number of visas, shown in Exhibit 4, which can be filtered by country and year. Additionally, Exhibit 4 is a bar graph with the average number of visas for each visa category which shows no significant difference between non-immigrant visas being issued between the Democratic or Republican party. From the information shown in the graph, the dominating visa category issued is the B-1/B-2 visas which allow persons who want to enter the US temporarily for business and for tourism.

Exhibit 3: Visa by Year and Political Party
https://prod-useast-b.online.tableau.com/#/site/zinaidadvoskina/workbooks/82521?:origin=card_share_link

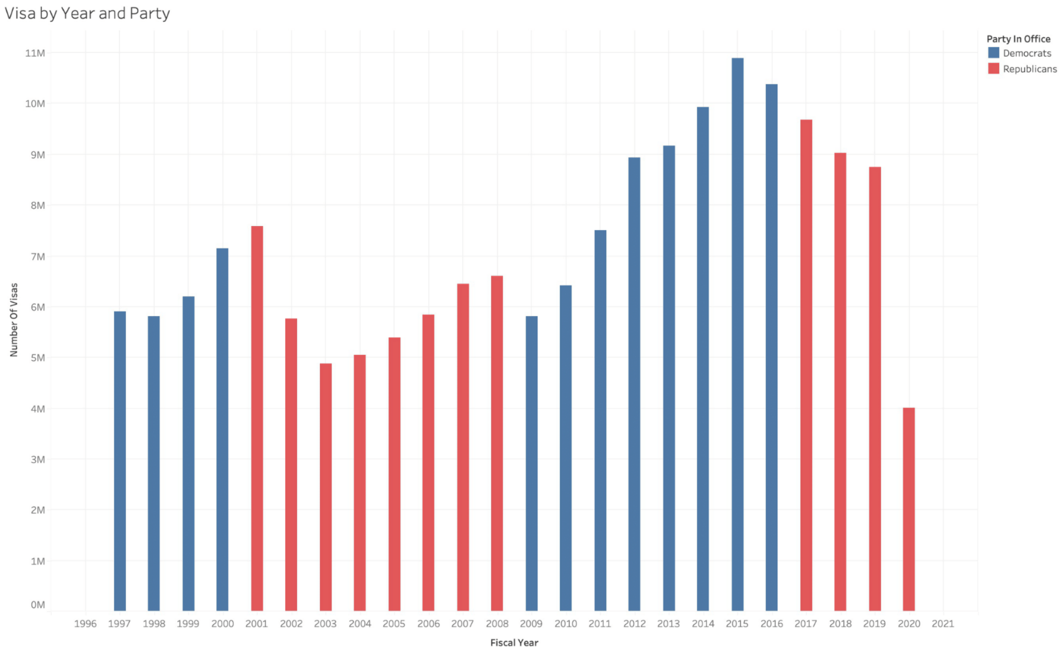


Exhibit 4: Visa Time Lapse
https://prod-useast-b.online.tableau.com/#/site/zinaidadvoskina/workbooks/82521?:origin=card_share_link

Visa Time Lapse - 2020

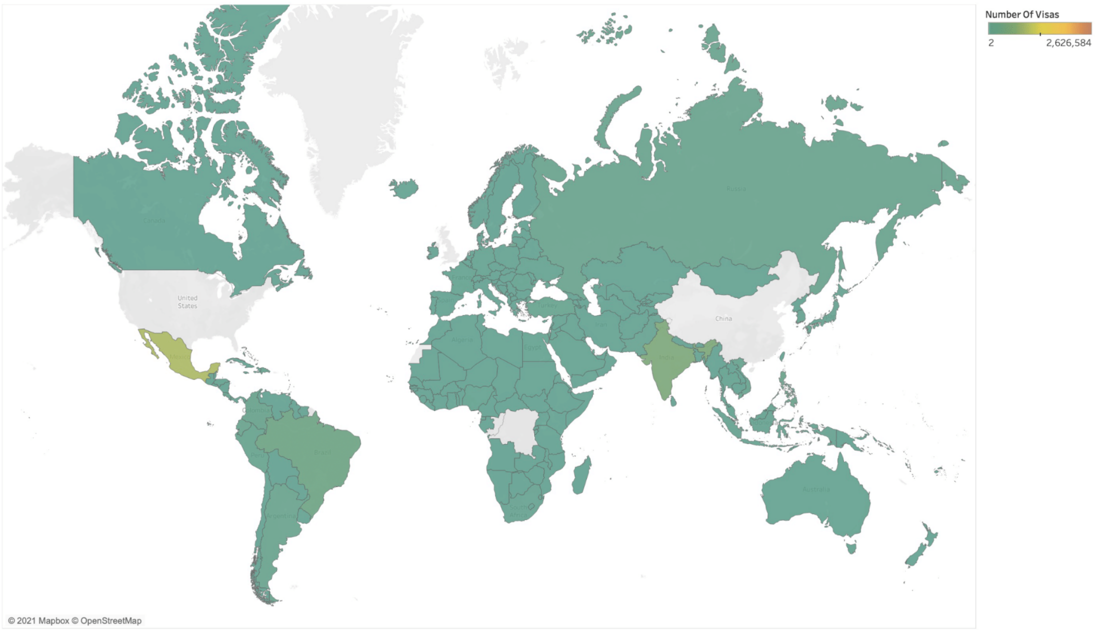
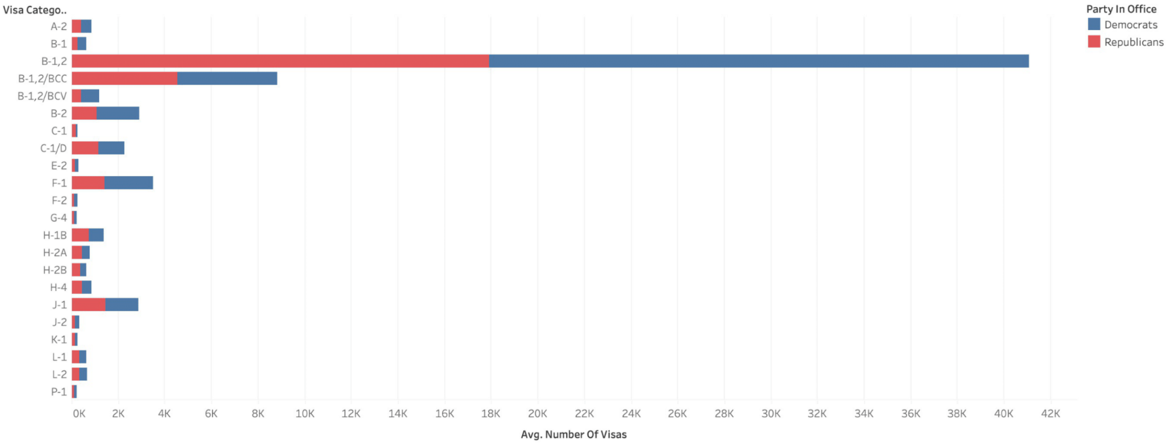


Exhibit 5: Visa Category Working
https://prod-useast-b.online.tableau.com/#/site/zinaidadvoskina/workbooks/82521?:origin=card_share_link

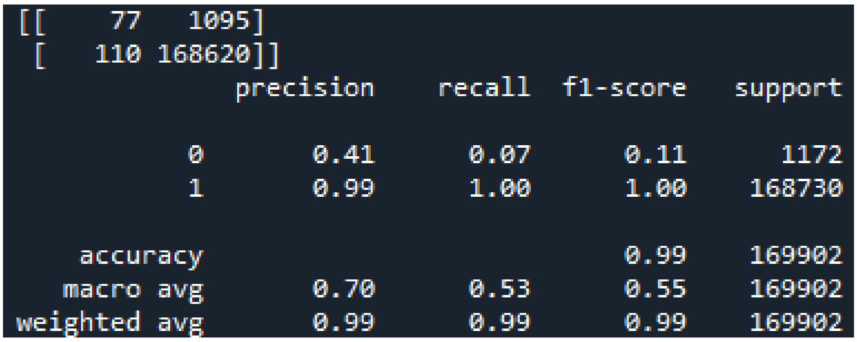
Visa Cat Working



- **KNN Model**

For our analysis we're using the KNN model for classification because it is the simplest and best-used approach with the following variables as predictors: Received month, Agent representing employer, Annual wage rate, Annual prevailing wage, PW wage level, H-1B dependent status, Support H1B status. This is a non-parametric model that serves as an advantage to our analysis as it allows more flexibility in our approach as it does not rely on calculating f(x). The Classification matrix in Exhibit 6 shows that the model performs very well predicting positive outcomes for visa approval, however it's accuracy is a lot smaller when predicting negative results. Overall, the model performs very well with f-1 score being 0.99. A potential disadvantage of using this model is that the distance measure becomes less reliable when there are irrelevant attributes.

Exhibit 6: KNN Model Classification Matrix



```
[[    77   1095]
 [   110 168620]]
              precision    recall  f1-score   support

           0       0.41      0.07      0.11      1172
           1       0.99      1.00      1.00    168730

    accuracy                           0.99    169902
   macro avg       0.70      0.53      0.55    169902
weighted avg       0.99      0.99      0.99    169902
```

- **Undersampling method for more accurate prediction models**

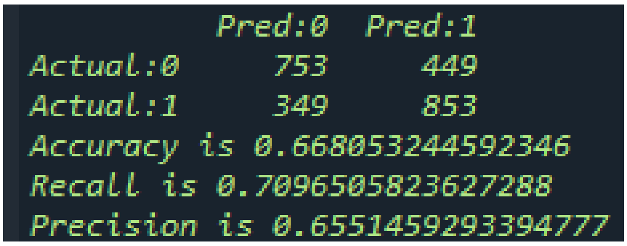
From what we can observe in our data, datasets are populated with approved results of visa applications. Almost 97% of data is positive, whereas negative Case Status percentage is very small. That results in highly biased prediction models towards positive outcomes, which means that we can not fully trust the model. What we can see in the KNN model in Exhibit 6, is that precision and recall for predicting negative results are very small, since the model is simply undertrained for making correct predictions.

One of the possible ways to solve such a problem is to implement undersampling of the dataset, prior to running any prediction models. Random underscaling will randomly eliminate data points from the dominating population and will even out the number of positive and negative outcomes for a more correct prediction.

In order to be able to run models with current equipment we have limited the dataset to 3 predictors which are Full Time Position, PW and New Employer. A larger number of predictors resulted in very long model processing times.

Our models are based on data from year 2020, however we constructed our code in such a general way that it can be used for any similar dataset from any other year or a combined dataset, but it will require much better equipment for a proper analysis.

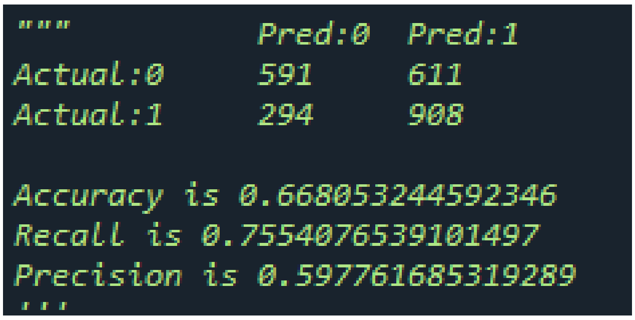
- **SVC model with undersampling method**



```
             Pred:0  Pred:1
Actual:0      753     449
Actual:1      349     853
Accuracy is 0.668053244592346
Recall is 0.7096505823627288
Precision is 0.6551459293394777
```

SVC model on an undersampled data set was first run with untuned parameters which result we can see on the image above. Model shows relatively high accuracy, recall and precision, which is more realistic compared to a highly biased model with no undersampling. We assume that we would be able to get better results for the model, however current equipment does not allow us to run models with more predictors. F1-score resulted to be 0.67 which is a fair model score to be considered, however it could be improved by tuning SVC parameters. We couldn’t calculate tuned parameters on our equipment due to memory issues and crashes, however, we implemented tuning in our code, so it would be possible to tune the model on better equipment.

  - **Logistic Regression model with undersampling method**

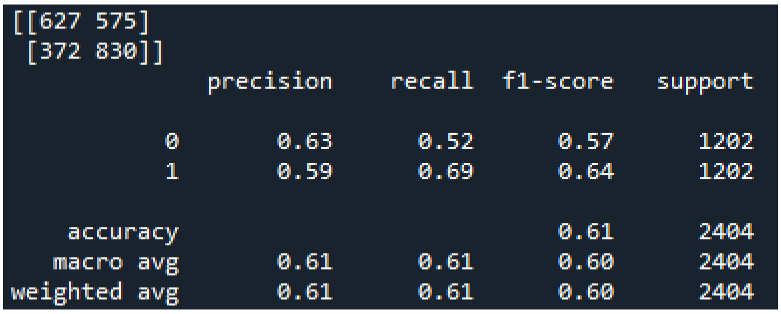


```
"""
             Pred:0  Pred:1
Actual:0     591     611
Actual:1     294     908

Accuracy is 0.668053244592346
Recall is 0.7554076539101497
Precision is 0.597761685319289
'''
```

Logistic regression model shows decent results and is more realistic on the same predictors than the initial KNN model with no undersampled data. Model performance can be further increased by tweaking predictors from the dataset, however we will still require better equipment to do so. Prediction model on visa decisions based on chosen predictors is fairly strong, however we can still see a great number of False Positive and False Negative results.

○ KNN model with undersampling method



```
[[627 575]
 [372 830]]
              precision    recall  f1-score   support

           0       0.63      0.52      0.57      1202
           1       0.59      0.69      0.64      1202

    accuracy                           0.61      2404
   macro avg       0.61      0.61      0.60      2404
weighted avg       0.61      0.61      0.60      2404
```

A new KNN model that we run on undersampled data, shows more realistic results, that are not biased towards a positive outcome. Model precision and recall is comparable to the ones we saw in other models, therefore we can say that chosen predictors have an impact on visa decision, however they only work in approximately 60% of situations. Further increase in the number of predictors may improve the model.

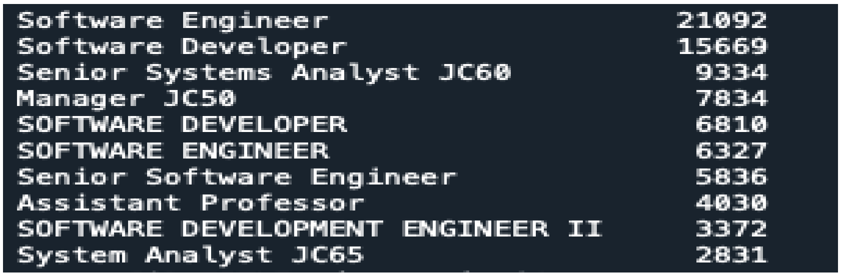
Part 5 - Recommendations/ Key Findings in 2020:

From our analysis, we can see in the exhibits below the key findings in 2020 for the top 10 results for job titles (Exhibit 7), companies (Exhibit 8), SOC code (Exhibit 9), city (Exhibit 10) and state (Exhibit 11). This information is valuable as non-immigrants can better understand the variables that can help them attain a working visa in the United States.

As a recommendation, the political party in office is not a great variable to determine the likelihood of an international student obtaining a professional working visa as there's no correlation shown on our results. In order to improve our prediction, it is necessary to include other variables that capture macro and micro factors.

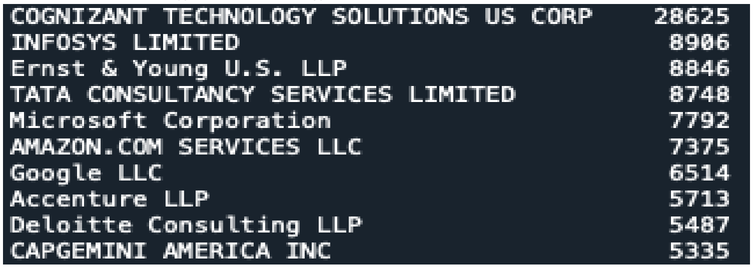
Moreover, models with different predictors show very high accuracy in predicting positive results of visa application, however they perform poorly in predicting negative results. This can be due to low amount of data on Visa denial cases that do not allow to increase the number of successful negative predictions

Exhibit 7: Top 10 Job Titles



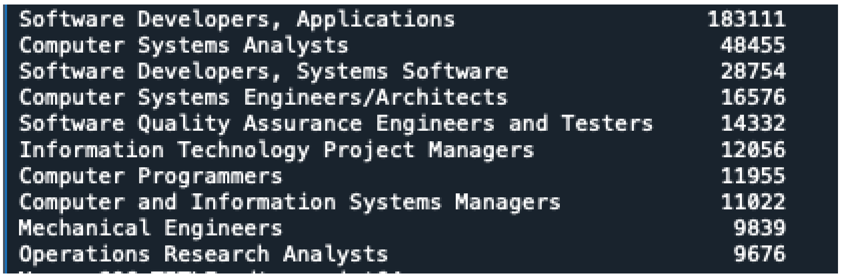
```
Software Engineer                      21092
Software Developer                     15669
Senior Systems Analyst JC60             9334
Manager JC50                            7834
SOFTWARE DEVELOPER                      6810
SOFTWARE ENGINEER                       6327
Senior Software Engineer                5836
Assistant Professor                     4030
SOFTWARE DEVELOPMENT ENGINEER II        3372
System Analyst JC65                     2831
```

Exhibit 8: Top 10 Companies



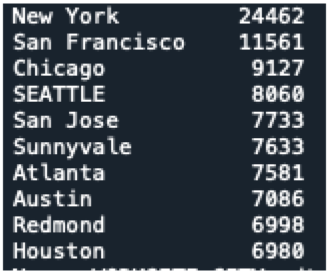
```
COGNIZANT TECHNOLOGY SOLUTIONS US CORP     28625
INFOSYS LIMITED                             8906
Ernst & Young U.S. LLP                      8846
TATA CONSULTANCY SERVICES LIMITED           8748
Microsoft Corporation                       7792
AMAZON.COM SERVICES LLC                     7375
Google LLC                                  6514
Accenture LLP                               5713
Deloitte Consulting LLP                     5487
CAPGEMINI AMERICA INC                       5335
```

Exhibit 9: SOC Code



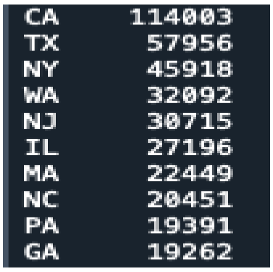
```
Software Developers, Applications                    183111
Computer Systems Analysts                             48455
Software Developers, Systems Software                 28754
Computer Systems Engineers/Architects                 16576
Software Quality Assurance Engineers and Testers      14332
Information Technology Project Managers               12056
Computer Programmers                                  11955
Computer and Information Systems Managers             11022
Mechanical Engineers                                   9839
Operations Research Analysts                           9676
```

Exhibit 10: City



```
New York         24462
San Francisco    11561
Chicago           9127
SEATTLE           8060
San Jose          7733
Sunnyvale         7633
Atlanta           7581
Austin            7086
Redmond           6998
Houston           6980
```

Exhibit 11: State



```
CA    114003
TX     57956
NY     45918
WA     32092
NJ     30715
IL     27196
MA     22449
NC     20451
PA     19391
GA     19262
```

Part 6 - Challenges and Future Work

______ Our primary challenge is due to the fact that our modeling and analysis require more sophisticated technological capacity. This is because the datasets that we are using are very large in size for the computers that our team members are using which has resulted in very long model processing times. As a result, it has been challenging to perform more in depth statistical analysis such as tuning our models due to speed, equipment, time and method restrictions.

For the future, we want to use the most up-to-date information for better accuracy that reflects real-life results as this can potentially be useful for departments that work with international students in colleges such as the Office of Global Services (OGS) and the co-op program at Northeastern University. However, immigration laws are constantly changing and data availability might be limited as it can be difficult to obtain demographic and immigration data for specific visa holders.

Based on our results, an interesting finding is that software engineers are at the top job title to obtain a working visa; however, they have the most denials; therefore, further modeling with appropriate technology can be done to look into it. With additional time and resources, such as a more sophisticated computer capacity, it could be possible to create interactive dashboards with complete historical database so that non-immigrant students can look into it when making choices before selecting a major in college or look into the companies where they could have a higher probability of obtaining a job after graduation, for example. This way, it can simplify the understanding of country-specific restrictions and requirements for different types of visas and determine better the impact of non-immigrant professionals in the U.S.

References:

Chen, C. (2019, May 7). *International students face challenges to work in U.S*. Michigan State University School of Journalism. https://news.jrn.msu.edu/2019/05/international-students-face-challenges-to-work-in-u-s/

USCIS. (2020, April 22). *Optional practical training (OPT) for F-1 students*. U.S. Citizenship and Immigration Services. https://www.uscis.gov/working-in-the-united-states/students-and-exchange-visitors/optional-practical-training-opt-for-f-1-students

USCIS. (2020, March 27). *H-1B specialty occupations, DOD cooperative research and development project workers, and fashion models*. U.S. Citizenship and Immigration Services. https://www.uscis.gov/working-in-the-united-states/temporary-workers/h-1b-specialty-occupations-dod-cooperative-research-and-development-project-workers-and-fashion