Predicting US Permanent Visa Application Decisions

Instructor: Prof Ying Lin

Team Members: Aneesh Phatak, Sankalp Singh

https://drive.google.com/file/d/1-1UhL8kOrvrhUAT5R37j88b_XJ22fyqz/view?usp=sharing

Problem Statement

- For a foreign citizen to work in the US, the employer must obtain a permanent labor certification and submit an immigration petition for that worker
- There are several immigration visas that lets a non-US citizen worker, work in the US including H-1B, F-1, J-1, L-1 visas
- Every year, a lot of tech giants like Google, Apple, Amazon etc. submit thousands of immigration petitions for their foreign workers
- Visa approval is a time-consuming process and there are times when the application process is denied causing a lot of problems for international workers, students and employers

Project Objective

- The objective of our project is to design a machine learning based system that could help in predicting US visa application decisions based on various input features like employer name, class of admission, country of citizenship etc.
- This system will be useful for the employers who actively file for visa applications for their international employees every year
- We will also try to identify what are the most important factors that help in predicting if a certain application will be certified or denied

Dataset Description

- The US Permanent Visa Applications dataset has been collected and distributed by US Department of Labor
- This dataset is available on Kaggle and contains a detailed information on 374k visa decisions between 2011-2016
- The target variable consists of 'Certified' and 'Denied' decisions for each application
- The dataset is highly imbalanced and consists of input features like employer name, employee education, class of admission etc.

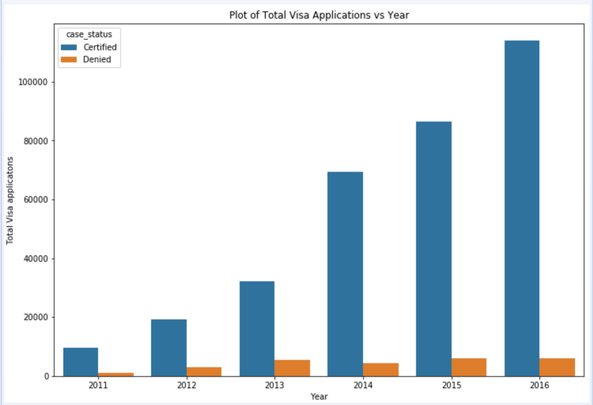
Methodology

- Here, we have followed the CRISP-DM methodology which is the cross industry standard data mining process
- This consists of following steps:
  - Business & Data Understanding
  - Data Pre-Processing and EDA
  - Model Building
  - Evaluation
  - Deployment

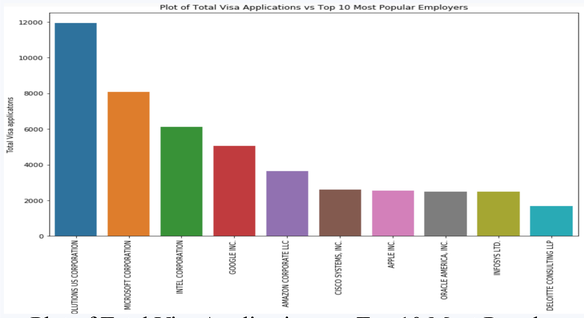
Data Cleaning and Pre-Processing

Techniques	Implementation
Transforming target attribute to a binary label	Combining the 'Certified-Expired' and 'Certified' instances into one category and removing 'Withdrawn' instances and just keeping 'Denied' instances
Feature Selection	Removing all the features containing more than 20% missing values as the dataset consists of 154 attributes
Removing Nan values	Imputing all the missing values with their respective mode values as most of the missing features are categorical
Under Sampling	Creating a sub-sample of the original dataset containing an equal number of instances for the binary target class labels as the original dataset is highly imbalanced
Label Encoding	Encoding all the features into numeric form to take care of categorical features
Standard Scaling	Scaling the range of all the input features so that each feature contribute to the final target label

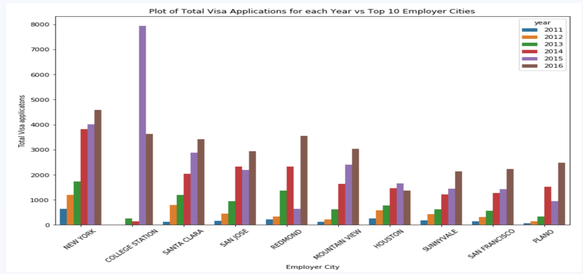
Exploratory Data Analysis



Plot of Total Visa Applications vs Year



Plot of Total Visa Applications vs Top 10 Most Popular Employers

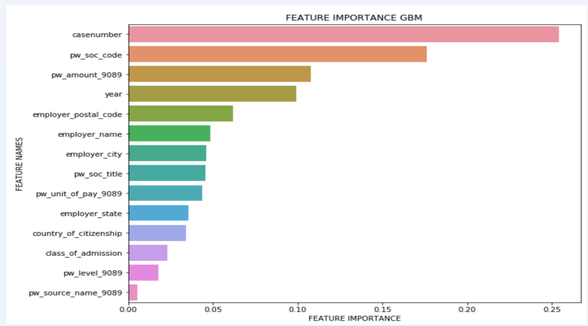


Plot of Total Visa Applications for each Year vs Top 10 Employer Cities

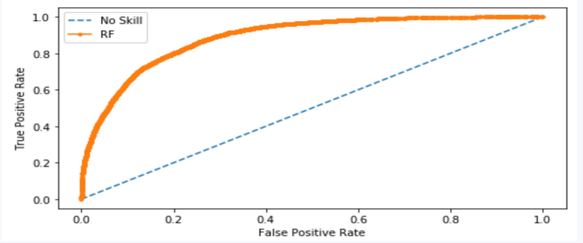
Model Building: ML Algorithms

Type	Algorithm	Evaluation Metrics
Classification	Logistic Regression	F1-Score
Classification	Random Forest	F1-Score
Classification	Gradient Boosting Machine	F1-Score
Classification	Linear Support Vector Machine	F1-Score
Classification	Artificial Neural Network	F1-Score

Feature Importance



ROC-AUC Curve



Evaluation

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	66	70	64	67	71.3
Random Forest	80	84	78	81	88.8
Gradient Boosting Machine	80	85	77	81	87.8
Linear SVM	66	70	64	67	-
ANN	69	68	70	70	75.5

Conclusion

- We have developed a system that can predict the US permanent visa application decisions. This system will help the employers by giving them insights on the factors that could lead to the approval or denial of a visa application
- Our system has also identified case number, year, employer name, employer city as some of the key features that are helpful in predicting the visa decisions

Future Scope

- In the future, we can try evaluating how some of the other techniques like oversampling and hybrid sampling can help in tackling the imbalanced dataset problem

References

Dataset Link:

https://www.kaggle.com/jboysen/us-perm-visas

https://towardsdatascience.com/imbalanced-data-in-classification-general-solution-case-study-169f2e18b017