EasyVisa

Data Analysis Observations & Recommendation

John D Noble

December 2021

Business Problem Overview & Solution Approach

This analysis is intended to use machine learning techniques to classify these applications and predict the "case status" or likelihood of being certified or no based on various factors.

Current State*

In FY 2016, the OFLC processed 776k applications for 1.7M positions. **A 9% increase from the previous year.**

- Business communities in the United States are facing high demand for human resources.
- A constant challenges is identifying and attracting the right talent, which is perhaps the most important element in remaining competitive.

The process of reviewing every case is becoming a tedious task as the number of applicants is increasing every year and is expensive!

What are we trying to solve for?

Can we leverage Machine Learning that can help in shortlisting the candidates having higher chances of VISA approval in a highly regulated environment?

Key Questions We Will Answer!

1. Do we have the proper data?
2. What modeling approaches will yield the best results?
3. Can we find the "right" candidates and reduce the HR resources required to do the initial screenings?

Financial Implications

While application volumes are continuing to rise 9% YoY, we need to reduce our total expense / we simply cannot hire people @ that trajectory to keep pace with the demand.

The Immigration and Nationality Act (INA) of the US permits foreign workers to come to the United States to work on either a temporary or permanent basis. The act also protects US workers against adverse impacts on their wages or working conditions by ensuring US employers' compliance with statutory requirements when they hire foreign workers to fill workforce shortages.

Data Description

Data Description

The data contains the different attributes of the employee and the employer. The detailed data dictionary is given below.

1. case_id: ID of each visa application
2. continent: Information of continent the employee
3. education_of_employee: Information of education of the employee
4. has_job_experience: Does the employee has any job experience? Y= Yes; N = No
5. requires_job_training: Does the employee require any job training? Y = Yes; N = No
6. no_of_employees: Number of employees in the employer's company
7. yr_of_estab: Year in which the employer's company was established
8. region_of_employment: Information of foreign worker's intended region of employment in the US.
9. prevailing_wage: Average wage paid to similarly employed workers in a specific occupation in the area of intended employment. The purpose of the prevailing wage is to ensure that the foreign worker is not underpaid compared to other workers offering the same or similar service in the same area of employment.
10. unit_of_wage: Unit of prevailing wage. Values include Hourly, Weekly, Monthly, and Yearly.
11. full_time_position: Is the position of work full-time? Y = Full-Time Position; N = Part-Time Position
12. case_status: Flag indicating if the Visa was certified or denied

Key Facts

We have a small data set for the analysis:

- 25k rows of data across these 12 variables
- The data consists of float64(1), int64(2), object(9)
- There are no missing or duplicate data
- None of the numerical data variables have any significant correlation

Data Preperation Made Before Modeling

1. # of employees does have 33 negative numbers…likely a capture error so we'll change those to be positive by taking the ABS value of the column “"no_of_employees"
2. Case ID is a system assigned sequence (highly cardinal) variable that adds no value … we'll drop this when we build the models.
3. We did check for outliers, and no adjustments are needed.
4. Mapped the word "certified" to a 1 and everything else 0 for the "target" we are trying to predict.

Data Description Details

The majority of these companies looking to fill positions have <6k employees. The average wage is ~74k.

Where are the candidates coming from?

1. Asia 16861
2. Europe 3732
3. North America 3292
4. South America 852
5. Africa 551
6. Oceania 19

Where does their educational background look like?

1. Bachelor's 10234
2. Master's 9634
3. High School 3420
4. Doctorate 2192

Have these applicants worked in the past?

1. Y 14802
2. N 10678

Do the open roles require prior experience?

1. N 22525
2. Y 2955

What region of the US is driving the demand?

1. Northeast 7195
2. South 7017
3. West 6586
4. Midwest 4307
5. Island 375

Salary and compensation profile?

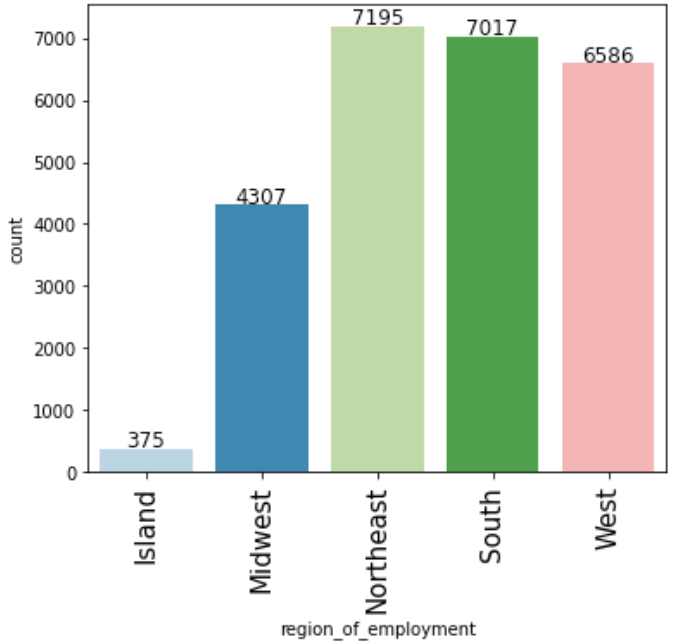
- 23k roles are full-time and offer yearly salaries

Historical acceptance rate of these candidates?

- *About ~67% of applicated are "certified'*

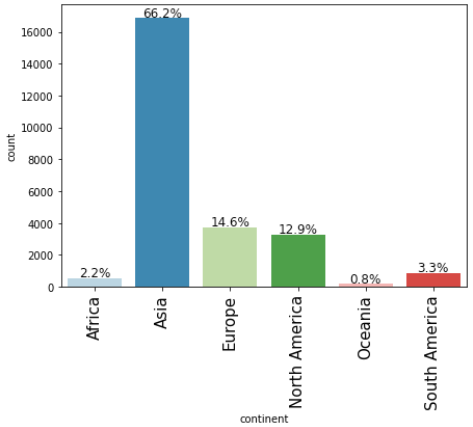
Exploratory Data Analysis (EDA): General Observations

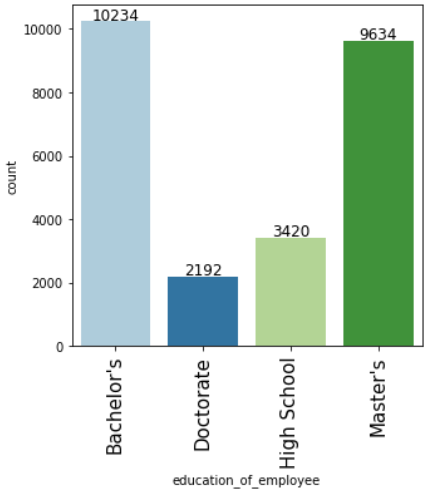
The demand is coming from the Northeast and South …



The supply is coming from

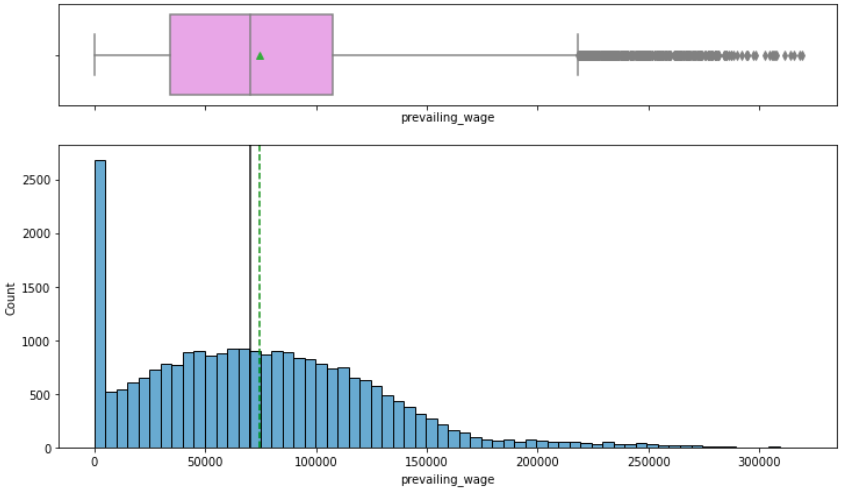
- Asia
- With a Bachelor's degree



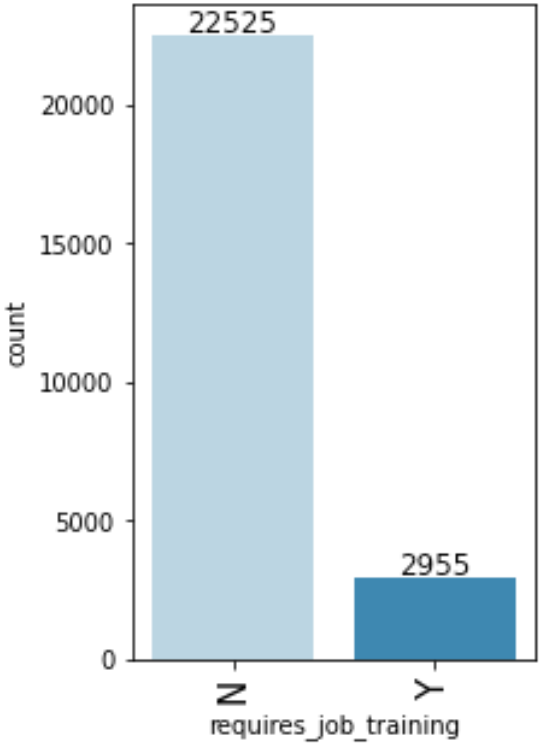


Exploratory Data Analysis (EDA): General Observations

The majority of roles are <$75k



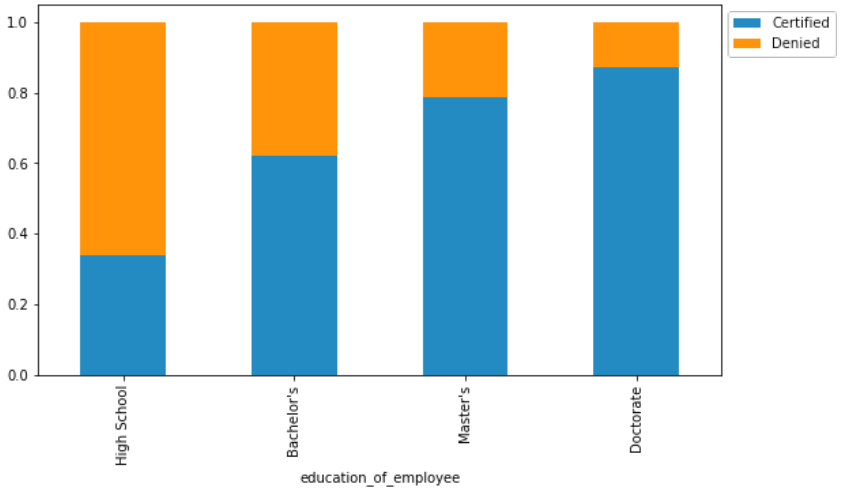
… and require no job training.

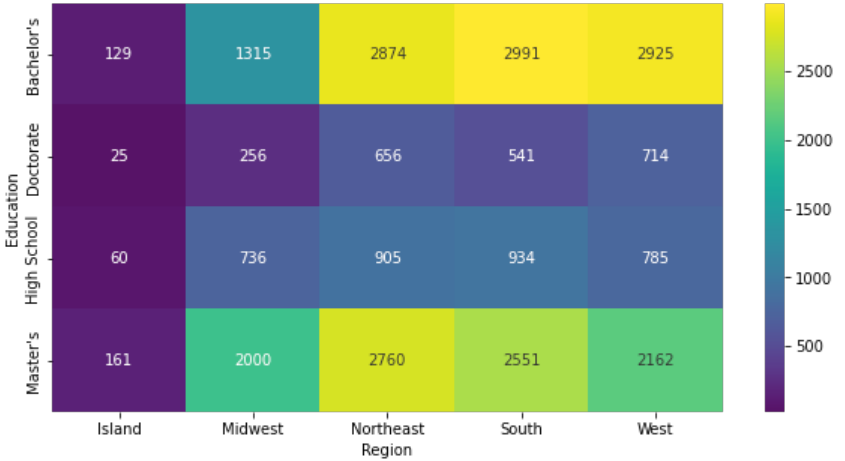


Exploratory Data Analysis (EDA): Detailed Observations

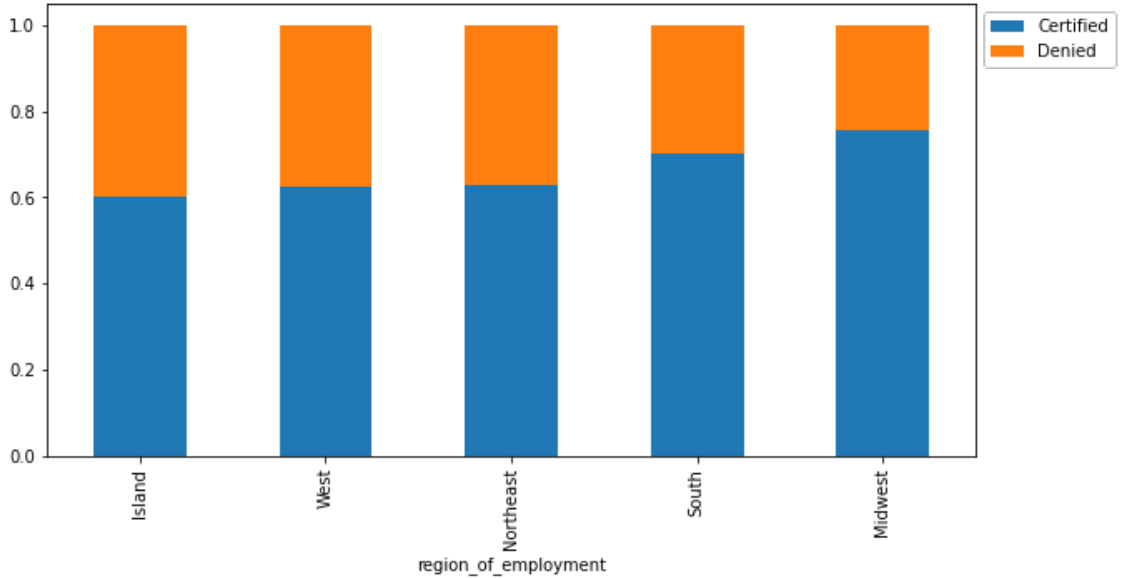
Master's/Doctorate acceptances were the highest given the specialized demand for certain roles where talent in the US simply doesn't exist enough to fill the vacancies.

… those candidates are going to jobs in the Northeast and South.



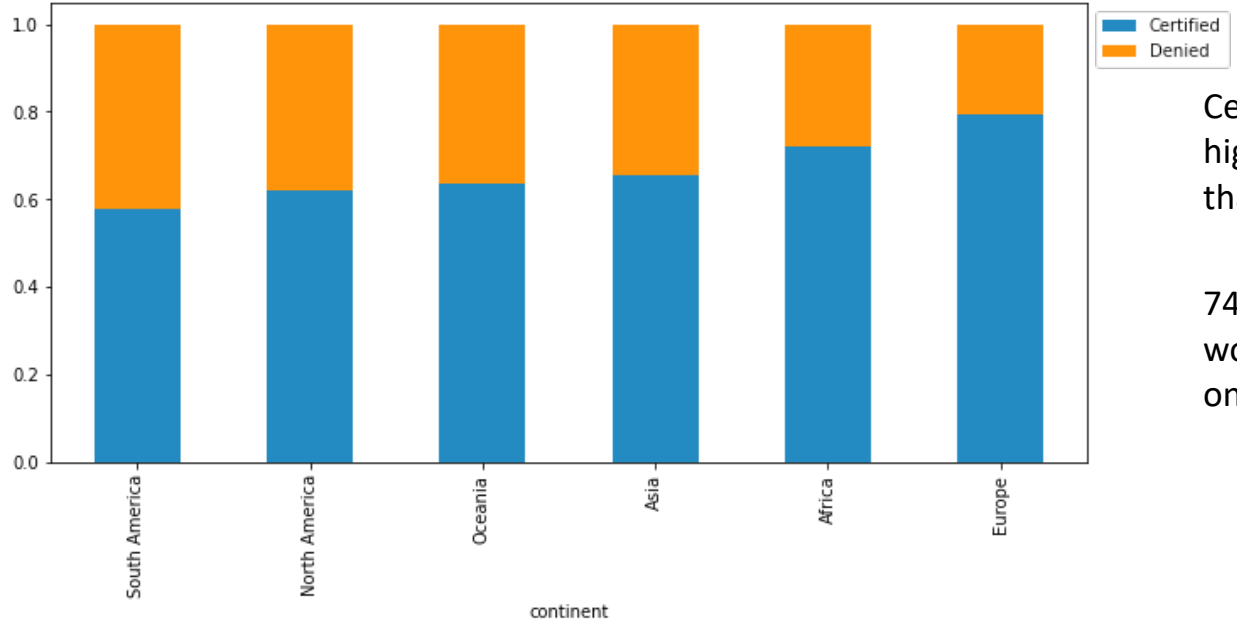


Exploratory Data Analysis (EDA): Detailed Observations



Acceptance by region, except in the Midwest, where it is more difficult to recruit taken who perceive the Northeast and South as more desirable locations to live and look for potential future work.

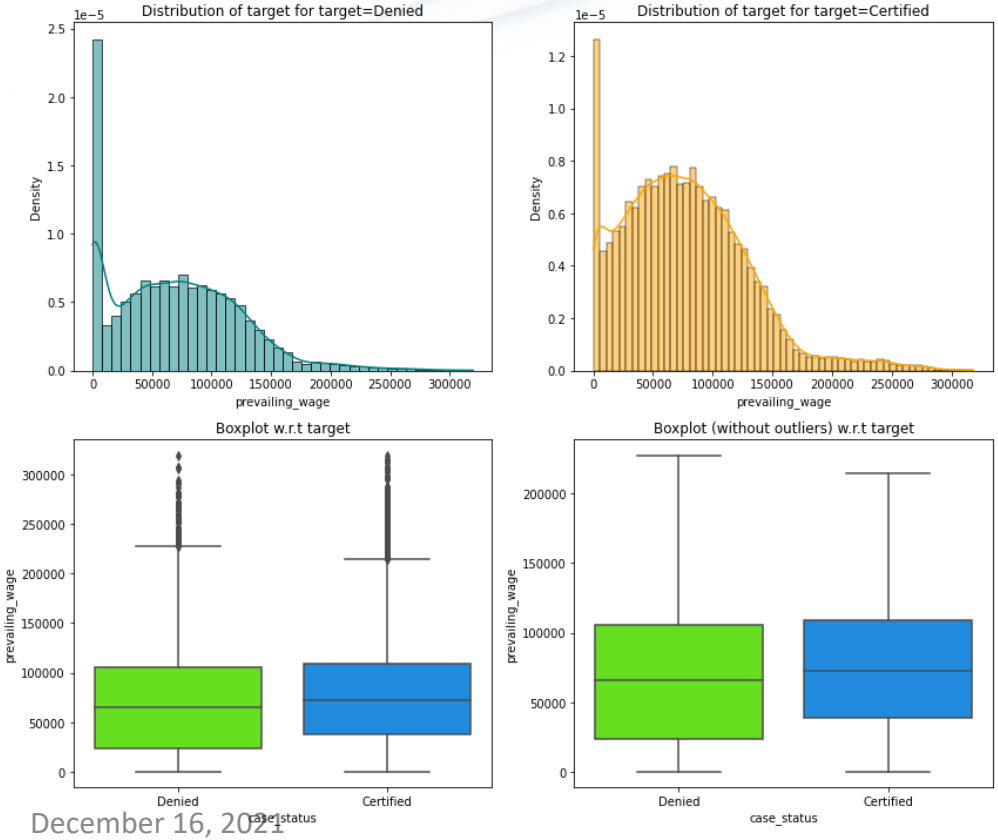
Exploratory Data Analysis (EDA): Detailed Observations



Certified applicants by region, are higher in Europe, Africa, and then Asia that are accepted.

74% of the resources hired, have prior work experience and will require some on the job training.

Exploratory Data Analysis (EDA): Detailed Observations

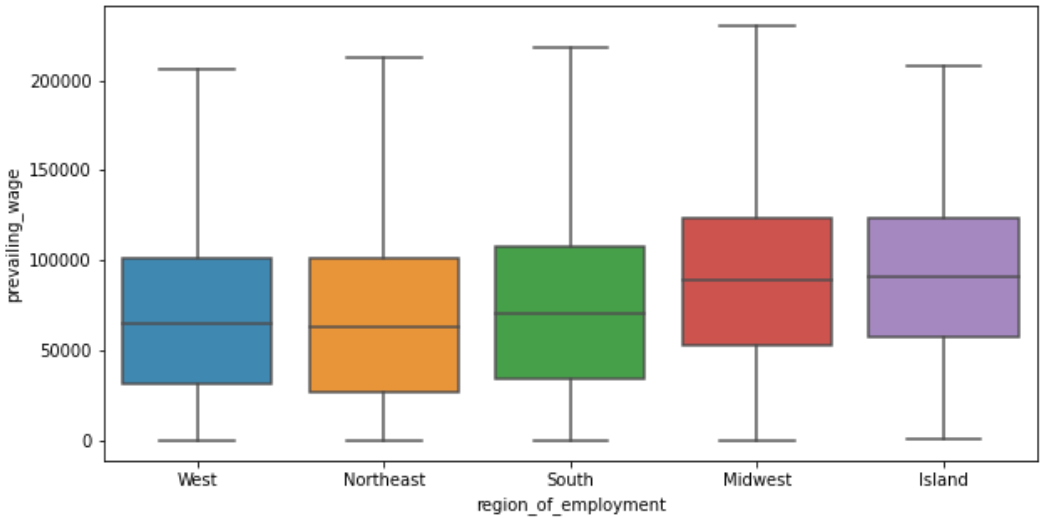


This is a highly regulated space, so the US government has established a prevailing wage to protect local talent and foreign workers.

Let's analyze the data and see if the visa status changes with the prevailing wage … **it has not had a significant impact on the prevailing total median wage compared to those denied vs. certified.**

Note: the Midwest and island regions have slightly higher median incomes given the prior slide and desirability to the location where employers have to offer higher incentives to get similar resources.

Exploratory Data Analysis (EDA): Detailed Observations



US labor demand does differ across regions. Where labor is widely available you see a closer relationship in the prevailing wage across the West, NorthEast, and South.

Where labor is less avaialble you can see that that the prevailing total median wage is higher in the Midwest and Islands. .

Note: the Midwest and island regions have slightly higher median incomes given the prior slide and desirability to the location where employers have to offer higher incentives to get similar resources.

CONSIDERATIONS FOR HOW TO APPROACH MODELING

Model can make wrong predictions as:

- Model predicts that the visa application will get certified but in reality, the visa application should get denied.
- Model predicts that the visa application will not get certified but in reality, the visa application should get certified.

Which case is more important?

- Both the cases are important as:
- If a visa is certified when it had to be denied a wrong employee will get the job position while US citizens will miss the opportunity to work on that position.
- If a visa is denied when it had to be certified the U.S. will lose a suitable human resource that can contribute to the economy.

How to reduce the losses?

- F1 Score can be used as the metric for evaluation of the model, greater the F1 score higher are the chances of minimizing False Negatives and False Positives.
- We will use balanced class weights so that model focuses equally on both classes.

Modeling Approaches

Can we improve the ~67% current acceptance rate using Machine Learning while reducing the number of human resources required to match future demand?

We will approach the testing with some well-known and common modeling techniques:

1. BaggingClassifier
2. RandomForestClassifier
3. AdaBoostClassifier
4. GradientBoostingClassifier
5. StackingClassifier

For our model we will use the standard SKLearn method for splitting the data:

```
Shape of Training set :  (17836, 21)
Percentage of classes in training set:
1    0.667919
0    0.332081
Name: case_status, dtype: float64
Shape of test set :  (7644, 21)
Percentage of classes in test set:
1    0.667844
0    0.332156
Name: case_status, dtype: float64
```

The data is balanced in terms of the classes represented in both the train and test sets which is important.

As a result we will use the F1 metric:

- Greater the F1 score higher are the chances of minimizing False Negatives and False Positives.
- We will use balanced class weights so that model focuses equally on both classes.

Model Performance Summary

Over our current acceptance rate, all of these models would show an improvement. So, any approach that introduces Machine Learning into the VISA determination process will have the desired impact of reducing the total HR headcount required to keep path with application growth.

TEST	Decision Tree	Tuned Decision Tree	Bagging Classifier	Tuned Bagging Classifier	Random Forest	Tuned Random Forest	Adaboost Classifier	Tuned Adaboost Classifier	Gradient Boost Classifier	Tuned Gradient Boost Classifier	XGBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	66.5%	70.7%	69.2%	72.4%	72.7%	73.8%	73.4%	71.7%	74.5%	74.3%	74.5%	74.4%	74.6%
Recall	74.3%	93.1%	76.4%	89.5%	84.7%	89.9%	88.5%	78.2%	87.6%	87.1%	87.7%	87.6%	87.6%
Precision	75.2%	71.5%	77.2%	74.4%	76.8%	75.5%	75.8%	79.2%	77.2%	77.3%	77.2%	77.2%	77.3%
F1	74.7%	80.9%	76.8%	81.3%	80.6%	82.1%	81.6%	78.7%	82.1%	81.9%	82.1%	82.1%	82.2%

Model Objective Recap

Which case is more important?

Both the cases are important as:

- If a visa is certified when it had to be denied a wrong employee will get the job position while US citizens will miss the opportunity to work on that position.
- If a visa is denied when it had to be certified the U.S. will lose a suitable human resource that can contribute to the economy.

How to reduce the losses?

- F1 Score can be used as the metric for evaluation of the model, greater the F1 score higher are the chances of minimizing False Negatives and False Positives.

Result Interpretation

1. The Stacking Classifier marginally beats out all the other methods; a consideration needs to made for explainability of both the model results and the model.
2. SIMPLE IS BETTER … MORE DATA REQUIRED and we'll continue to run A/B testing to further refine the modeling approach.
    - There is a case to be made for "simple" and use the Tuned RF as it is just easier to explain.

Model Performance Detail (Train vs. Test)

Over our current acceptance rate, all of these models would show an improvement. So, any approach that introduces Machine Learning into the VISA determination process will have the desired impact of reducing the total HR headcount required to keep path with application growth.

TRAIN	Decision Tree	Tuned Decision Tree	Bagging Classifier	Tuned Bagging Classifier	Random Forest	Tuned Random Forest	Adaboost Classifier	Tuned Adaboost Classifier	Gradient Boost Classifier	Tuned Gradient Boost Classifier	XGBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	100%	71%	99%	100%	100%	77%	74%	72%	76%	76%	76%	76%	77%
Recall	100%	93%	99%	100%	100%	92%	89%	78%	88%	88%	88%	88%	89%
Precision	100%	72%	99%	99%	100%	78%	76%	79%	78%	79%	78%	78%	79%
F1	100%	81%	99%	100%	100%	84%	82%	79%	83%	83%	83%	83%	84%

- One can clearly see the Decision Trees (as expected) overfit the training data
- Compared to the boosting methods

TEST	Decision Tree	Tuned Decision Tree	Bagging Classifier	Tuned Bagging Classifier	Random Forest	Tuned Random Forest	Adaboost Classifier	Tuned Adaboost Classifier	Gradient Boost Classifier	Tuned Gradient Boost Classifier	XGBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	66.5%	70.7%	69.2%	72.4%	72.7%	73.8%	73.4%	71.7%	74.5%	74.3%	74.5%	74.4%	74.6%
Recall	74.3%	93.1%	76.4%	89.5%	84.7%	89.9%	88.5%	78.2%	87.6%	87.1%	87.7%	87.6%	87.6%
Precision	75.2%	71.5%	77.2%	74.4%	76.8%	75.5%	75.8%	79.2%	77.2%	77.3%	77.2%	77.2%	77.3%
F1	74.7%	80.9%	76.8%	81.3%	80.6%	82.1%	81.6%	78.7%	82.1%	81.9%	82.1%	82.1%	82.2%

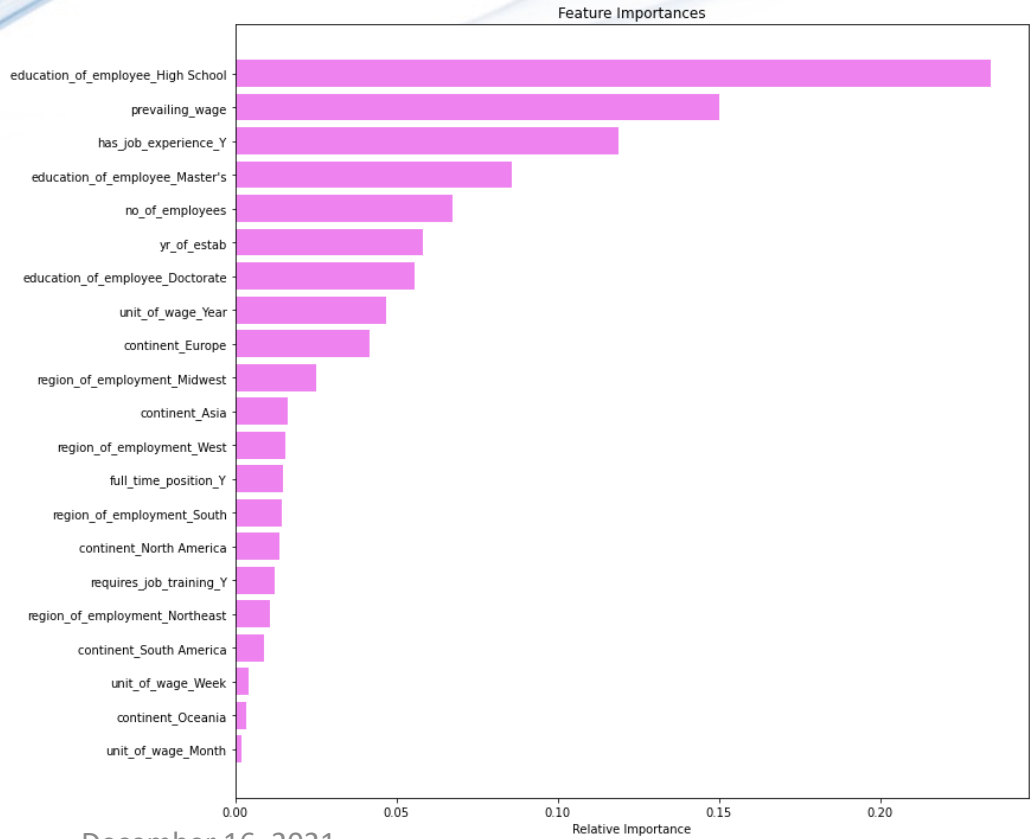
- Boosting methods with Hyperparamer Tuning & Stacking Classifier Works Best!

Model Performance Summary

Hyperparameter Tuning – we found the optimal combination for each model of the parameters using GridSearchCv method, and then finally a Stacking Classifier to combine the 'best-of-the-optimal' tuned models, and then use XGBoost to get the final prediction.

1. DecisionTreeClassifier(class_weight='balanced', max_depth=10, max_leaf_nodes=2, min_impurity_decrease=0.0001, min_samples_leaf=3, random_state=1)
2. BaggingClassifier(max_features=0.7, max_samples=0.7, n_estimators=100, random_state=1)
3. RandomForestClassifier(max_depth=10, max_features='sqrt', min_samples_split=7, n_estimators=20, oob_score=True, random_state=1)
4. AdaBoostClassifier(base_estimator=DecisionTreeClassifier(class_weight='balanced', max_depth=1, random_state=1),
5. GradientBoostingClassifier(init=AdaBoostClassifier(random_state=1), max_features=0.8, n_estimators=200, random_state=1, subsample=1)
6. XGBClassifier(colsample_bylevel=0.9, colsample_bytree=0.9, eval_metric='logloss', gamma=5, n_estimators=200, random_state=1)
7. StackingClassifier(estimators=[('AdaBoost', AdaBoostClassifier(random_state=1)),
    a. ('Gradient Boosting',GradientBoostingClassifier(init=AdaBoostClassifier(random_state=1), max_features=0.8, n_estimators=200, random_state=1, subsample=1)),
    b. ('Random Forest', RandomForestClassifier(max_depth=10, max_features='sqrt', min_samples_split=7, n_estimators=20, oob_score=True, random_state=1))],
    c. final_estimator=XGBClassifier(colsample_bylevel=0.9, colsample_bytree=0.9, eval_metric='logloss', gamma=5, n_estimators=200, random_state=1))

Model Feature Importance



Result Interpretation

This graph shows the top features when it comes to the Stacking Classifier and the impact of those features on successfully predicating whether a VISA will or will no be granted.

Whether an individual has a "high school" education is the most significant feature, followed by 'prevailing wage', then 'prior experience' in determining whether they get certified or not.

Business Insights and Recommendations

1. We can use this predictive model for any Visa applicant assuming we know the data about them in advance and with that information predict the likelihood of getting certified or not.
2. Additionally, we can use the model and compare that against future labor demand demand to “look back” and determine whether the pipeline of applicants is sufficient to meet that demand; if not we can proactively anticipate avg median wages increases and monitor that against the benchmarks established by the protects US workers against adverse impacts
3. What are the key factors that affect the target variable?
    1. High School Education
    2. Prevailing Wage
    3. Prior Job Experience
    4. Masters Certification
    5. No of Employees
    6. Other Important features - Midwest, Doctorate, Europe (not in order)
4. Compared to the current ~67% certification rate, we do not know the current rate of the False Negatives and False Positives. We know based on the results that we can achieve a proper classification to ~82%+ of the time

Business Insights and Recommendations

5. Implementing a machine learning model, will eliminate the need to higher future resources to review the applications to keep up with demand.
    a. It is also very likely that we can redeploy current resources as a result of the efficiency gain (automation) by reducing or eliminating the risks those False Negatives/Positives for certain job categories
    b. We'd also be able to determine which employers are sending the greatest number applications and what is the ratio of certifications by job sector.
    c. At the same time, figure out the most common job title and their relative wages for which the greatest number of applications are filed. .
6. Additionally, we can use the model and compare that against future labor demand demand to "look back" and determine whether the pipeline of applicants is sufficient to meet that demand; if not we can proactively suggest avg median wages increases and monitor that against the benchmarks established by the protects US workers against adverse impacts
    a. Employers in regions like the Midwest where demand can have an impact on avg median wages, employers can "subscribe" to this service for a "fee" to both measure and adjust their incentives to ensure they can meet the demand.
7. "Year_established" – there are a number of organizations prior to 1976. While we do not know based on the data (no industry subsegment), however if we think about this period in US history, these jobs may be manufacturing and/or service (hotels, etc). This demand is certainly not in the "high-tech" space in which we'd expect to see higher wage inflation; an unintended cost of keep avg median income lower on average because non-US resources are willing to take those roles where US-based resources are not.

Exploratory Data Analysis (EDA): Detailed Appendix

Dataset statistics

Number of variables	12
Number of observations	25480
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	2.3 MiB
Average record size in memory	96.0 B

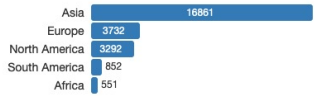
Variable types

Categorical	6
Boolean	3
Numeric	3

Exploratory Data Analysis (EDA): Detailed Appendix

continent
Categorical

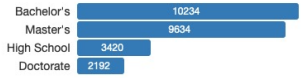
Distinct	6
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	199.2 KiB



Toggle details

education_of_employee
Categorical

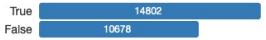
Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	199.2 KiB



Toggle details

has_job_experience
Boolean

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	25.0 KiB



Toggle details

requires_job_training
Boolean

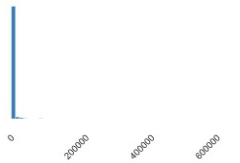
Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	25.0 KiB

Toggle details

Exploratory Data Analysis (EDA): Detailed Appendix

no_of_employees
Real number (ℝ)

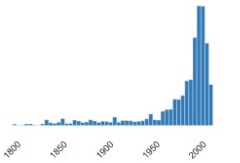
Distinct	7105	**Minimum**	-26
Distinct (%)	27.9%	**Maximum**	602069
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	33
Infinite (%)	0.0%	**Negative (%)**	0.1%
Mean	5667.04321	**Memory size**	199.2 KiB



Toggle details

yr_of_estab
Real number ($\mathbb{R}_{\geq 0}$)

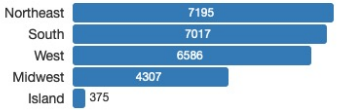
Distinct	199	**Minimum**	1800
Distinct (%)	0.8%	**Maximum**	2016
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1979.409929	**Memory size**	199.2 KiB



Toggle details

region_of_employment
Categorical

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	199.2 KiB

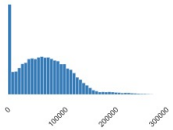


Toggle details

Exploratory Data Analysis (EDA): Detailed Appendix

prevailing_wage
Real number (ℝ≥0)

Distinct	25454	**Minimum**	2.1367
Distinct (%)	99.9%	**Maximum**	319210.27
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	74455.81459	**Memory size**	199.2 KiB



Toggle details

unit_of_wage
Categorical

Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	199.2 KiB

Toggle details

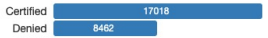
full_time_position
Boolean

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	25.0 KiB

Toggle details

case_status
Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	199.2 KiB



Toggle details

< TARGET VARIABLE

Exploratory Data Analysis (EDA): Detailed Appendix Sample Data

	CASE_ID	CONTINENT	EDUCATION_OF_EMPLOYEE	HAS_JOB_EXPERIENCE	REQUIRES_JOB_TRAINING	NO_OF_EMPLOYEES	YR_OF_ESTAB	REGION_OF_EMPLOYMENT	PREVAILING_WAGE	UNIT_OF_WAGE	FULL_TIME_POSITION	CASE_STATUS
0	EZYV01	Asia	High School	N	N	14513	2007	West	592.20	Hour	Y	Denied
1	EZYV02	Asia	Master's	Y	N	2412	2002	Northeast	83,425.65	Year	Y	Certified
2	EZYV03	Asia	Bachelor's	N	Y	44444	2008	West	122,996.86	Year	Y	Denied
3	EZYV04	Asia	Bachelor's	N	N	98	1897	West	83,434.03	Year	Y	Denied
4	EZYV05	Africa	Master's	Y	N	1082	2005	South	149,907.39	Year	Y	Certified
5	EZYV06	Asia	Master's	Y	N	2339	2012	South	78,252.14	Year	Y	Certified
6	EZYV07	Asia	Bachelor's	N	N	4985	1994	South	53,635.39	Year	Y	Certified
7	EZYV08	North America	Bachelor's	Y	N	3035	1924	West	418.23	Hour	Y	Denied
8	EZYV09	Asia	Bachelor's	N	N	4810	2012	Midwest	74,362.19	Year	Y	Certified
9	EZYV10	Europe	Doctorate	Y	N	2251	1995	South	67,514.76	Year	Y	Certified