EasyVisa - Project

Data driven solution to facilitate visa approvals for OFLC

Contents

Business Problem Overview and Solution Approach

- ## Context of Business Problem

  With growing demand of business communities in United States are facing high demand for human resources, and with Immigration and Nationality act (INA) of US permits 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. The immigration programs are administered by the Office of Foreign Labor Certification (OFLC). With every passing year the number of applications are increasing.

- ## Problem to tackle

  FY 2016, the OFLC processed 775,979 employer applications for 1,699,957 positions for temporary and permanent labor certifications. This was a nine percent increase in the overall number of processed applications from the previous year. The process of reviewing every case is becoming a tedious task as the number of applicants is increasing every year. The increasing number of applicants every year calls for a Machine Learning based solution that can help in shortlisting the candidates having higher chances of VISA approval.

- ## Financial implications

  The manual review of each application –
  1. Takes significant amount of time and human resource.
  2. Manual review process could be prone to error if the workload will be huge, erroneous certification could lead to legal implication.
  3. Both of this will lead to added cost for OFLC and lead to financial implications.

- ## How to use ML model to solve the problem

  With a good robust predictive model where the model can predict suitable profiles for applicants for whom visa should be certified or denied, will facilitate process these growing number of applications timely manner, beneficial for OFLC saving cost on time and human resources and also helps business communities get required foreign resources on time.

Data Overview

- Data Dictionary

Variables	Description
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	Prior job experience of the employee, Y – Yes, N - No
5. requires_job_training	Employee require Job training or not, Y – Yes, N - No
6. no_of_employees	No of employees in the employer's company
7. yr_of_estab	Year on which employer's company was established
8. region_of_employment	Intended region of employment in US
9. prevailing_wage	Prevailing wage for the employee
10. unit_of_wage	Unit of prevailing wage
11. full_time_position	Position is fulltime or not. Y – Yes, N - No
12. case_status	Case status as flag value indicating certified/denied

Observations	Variables
25480	12

Notes:

1. Dataset looked consistent with the dictionary provided.
2. There are no duplicate values
3. There are no null/missing values across all variables.

Missing value counts	
case_id	0
continent	0
education_of_employee	0
has_job_experience	0
requires_job_training	0
no_of_employees	0
yr_of_estab	0
region_of_employment	0
prevailing_wage	0
unit_of_wage	0
full_time_position	0
case_status	0

Duplicate value counts
0

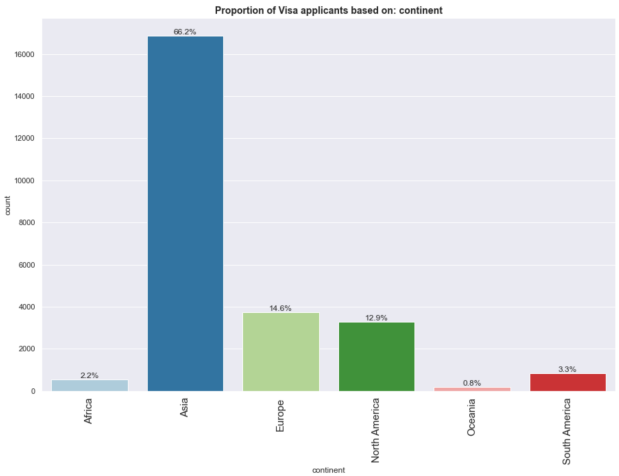
Data Overview contd..

- Brief description of significant manipulations made to raw data

1. Negative values in no of employee variable are just taken as absolute values.
2. Outliers' treatment was not done as the models we have built are not sensitive to outliers.
3. Spaces and special characters were removed from the values present in continent and education of employee variables
4. We have extracted two new features employer size(based on no of employees binned into groups) and prevailing wage group(based on prevailing wage binned into groups)
5. For model building we have dropped case_id variable as it’s just a unique identifier of each case and will not add any value to the predictive model, the target variable case_status converted to numeric target variable with certified as 1 and denied as 0.

Exploratory Data Analysis – Univariate Analysis

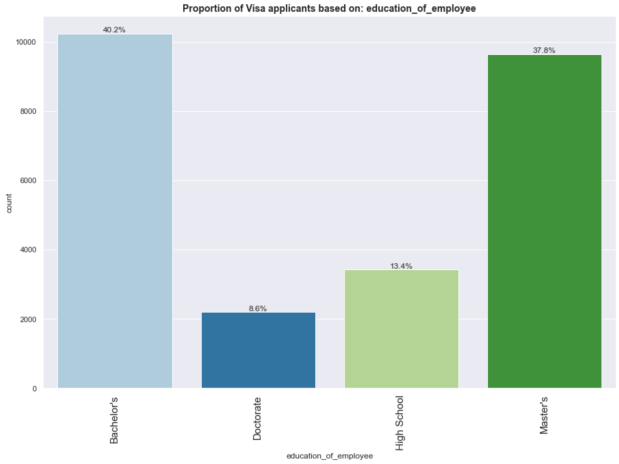
Proportion of continent



Notes:

1. With ~66% Asia has the highest number of applicants.
2. Oceania has the lowest number of applicants, only ~.8%
3. Europe and North America shows ~14% and ~13% of applicants in the dataset

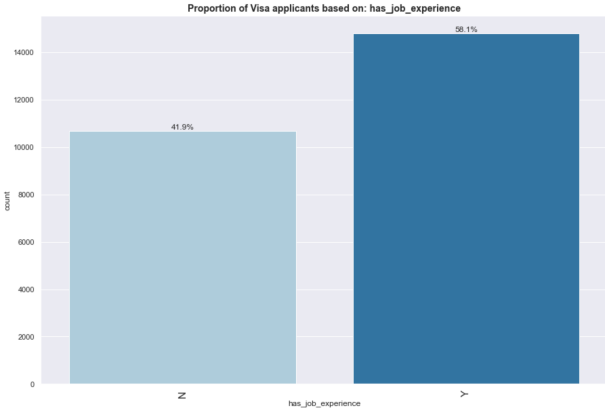
Proportion of education of employee



Notes:

1. Majority of applicants are from two education groups Bachelor's and Master's
2. ~40% applicants are holding Bachelor's degree, and it's highest in education group.
3. Only ~8% of applicants have completed Doctorate and lowest among the education group

Proportion of prior job experience

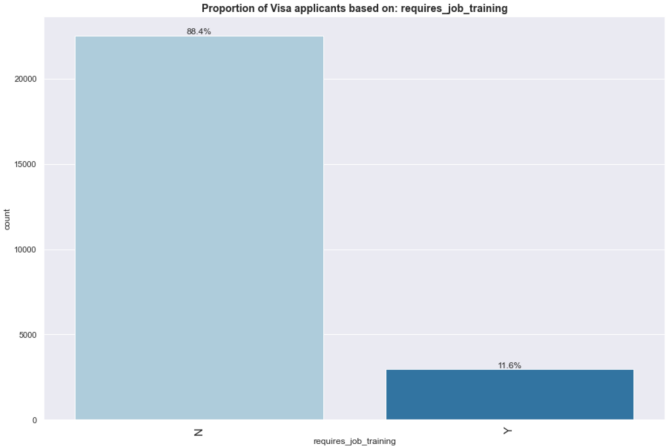


Notes:

1. ~58% of applicants has prior job experience
2. ~42% doesn't have prior job experience

Exploratory Data Analysis – Univariate Analysis

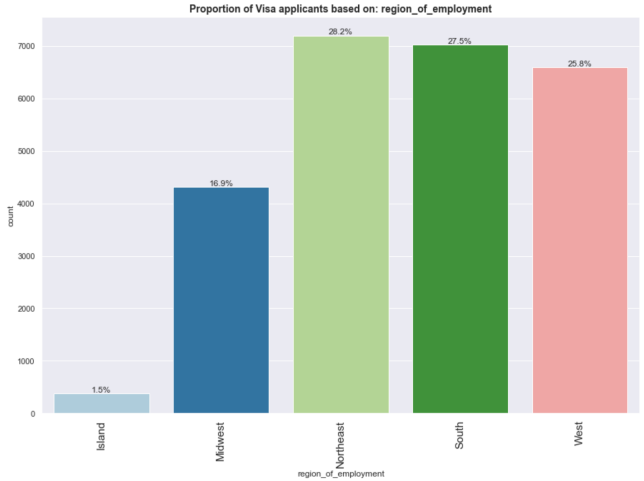
Proportion of requires job training



Notes:

1. Predominantly applicants doesn't require job training.
2. ~88% of applicants doesn't require any job training, only ~12% of applicants who requires training.

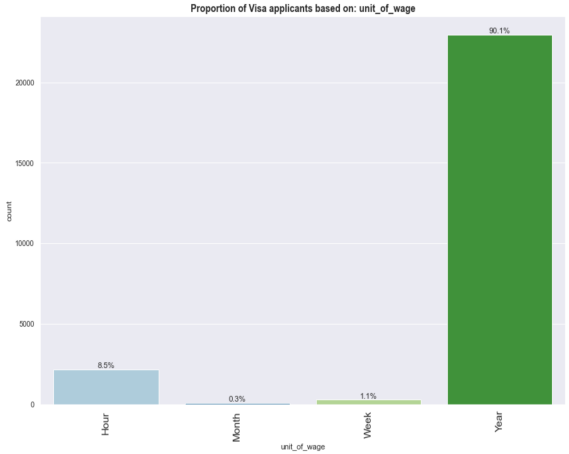
Proportion of region of employment



Notes:

1. Intended employment region Northeast, South and West all shows almost similar number of applicants, in range of ~28-~25% in each region.
2. US Island shows the lowest ~1.5% of applicants, and Midwest shows ~16%

Proportions of unit of wage

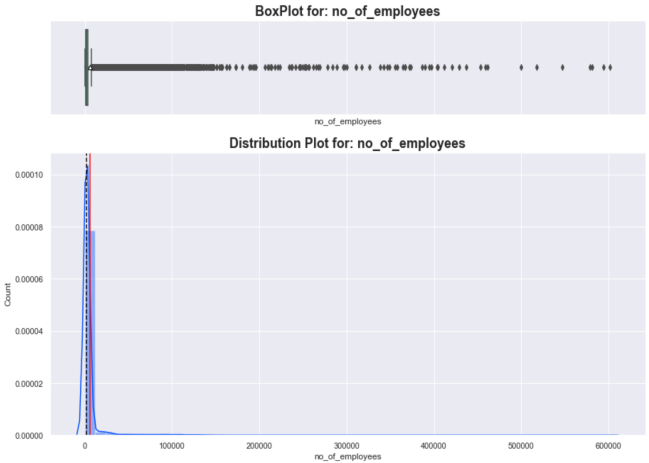


Notes:

1. ~90% of applicants wage is Yearly wage.
2. Second highest is hourly wage ~8%

Exploratory Data Analysis – Univariate Analysis

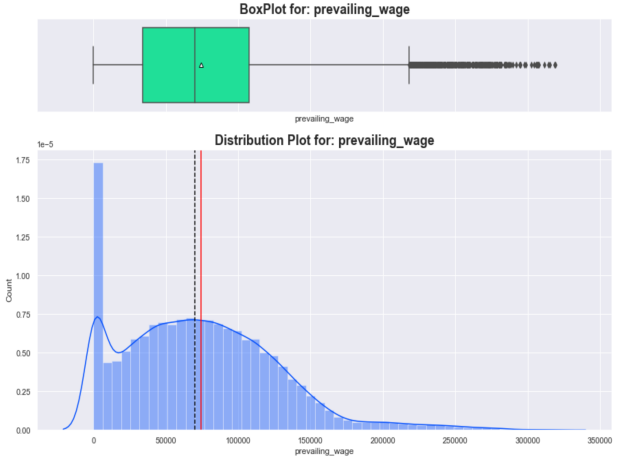
Distribution of no of employees



Notes:

1. No of employees shows as high as ~600K for some employees.
2. Majority of employers have average employee strength of ~5K or less

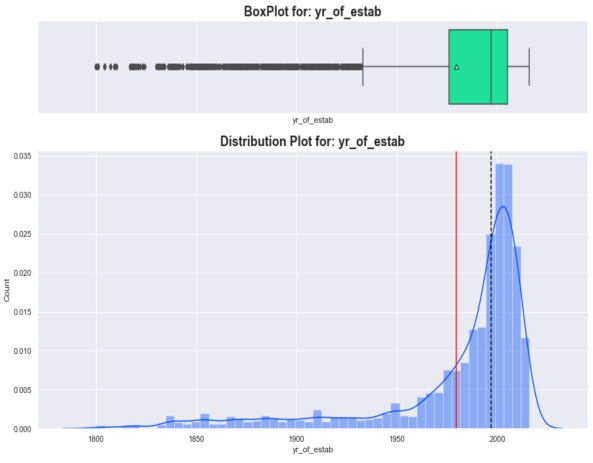
Distribution of prevailing wage



Notes:

1. Prevailing wage shows as high as $350K as wage level
2. There is a peak below $5K as prevailing wage, due to unit difference and as we saw we have significant applicants of hourly wage.

Distribution of year of establishment

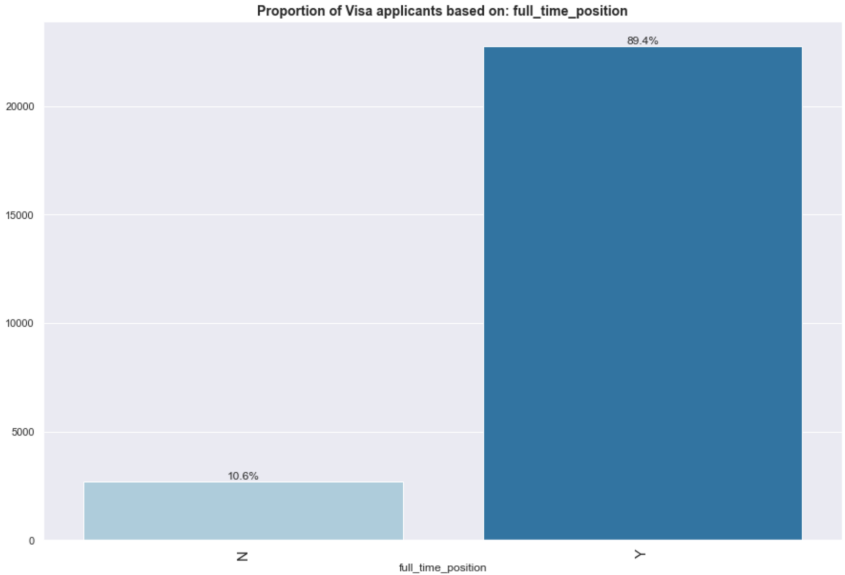


Notes:

1. Data is heavily left skewed, with employers established as early as 1800.
2. Majority of employers are established ~1975 or after.

Exploratory Data Analysis – Univariate Analysis

Market Segment proportions

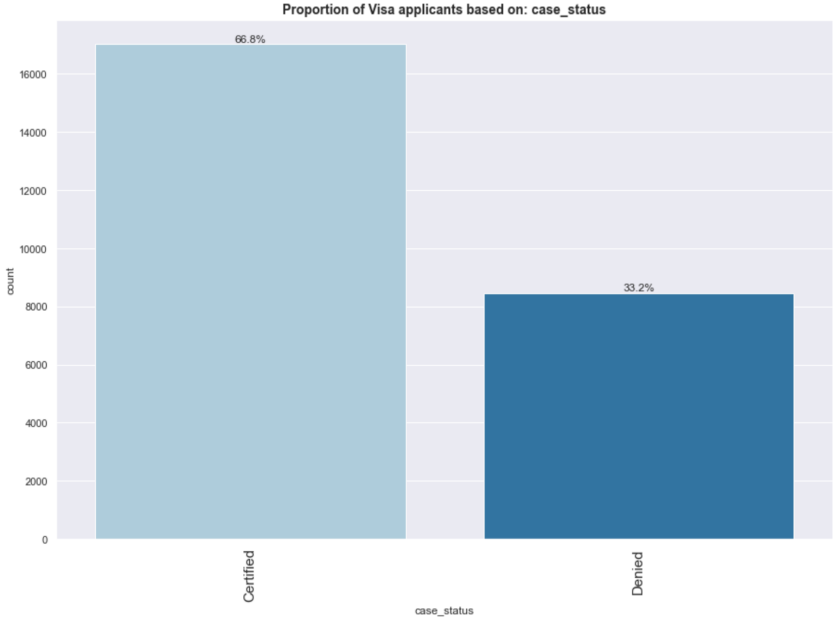


Notes:

1. ~89% of applicants have applied for full time positions
2. Only ~11% of applicants are for part time/other positions

Repeated guest proportions

Previous Cancellations



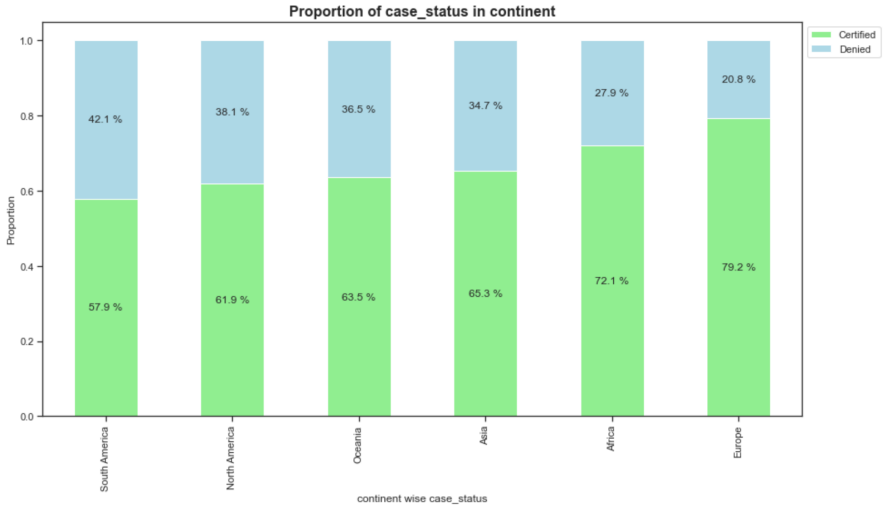
Notes:

1. ~67% of applications are certified
2. ~33% applications are denied

Previous Not Cancelled Bookings

Exploratory Data Analysis – Bivariate Analysis

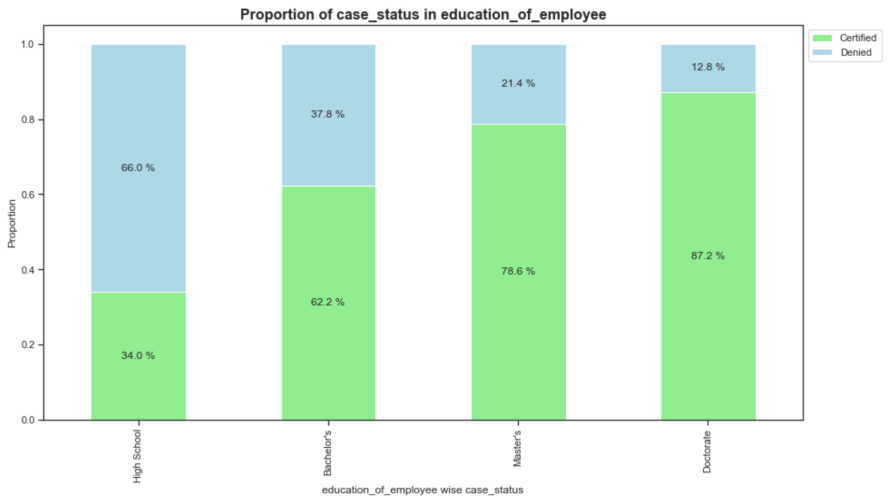
Case status per continent



Notes:

1. Europe has the highest visa certified rate, close to ~79% of applicants have their visa certified, ~21% got rejected
2. Africa has the second highest approval rate - ~72%
3. South America shows the lowest approval rate with ~58%
4. Asia, Oceania, North America all has their approval rate ranging ~62% to ~65%

Case status per education group

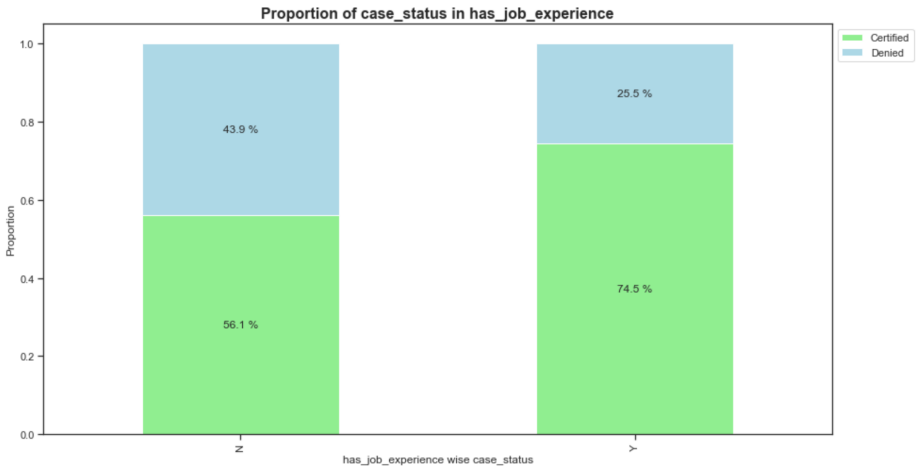


Notes:

1. With higher degree the chances of visa being certified increases.
2. Doctorate, the highest form of degree in the dataset shows ~87% of approval rate.
3. Applicants with only high school degree has approval rate of ~34% which is the lowest

Exploratory Data Analysis – Bivariate Analysis

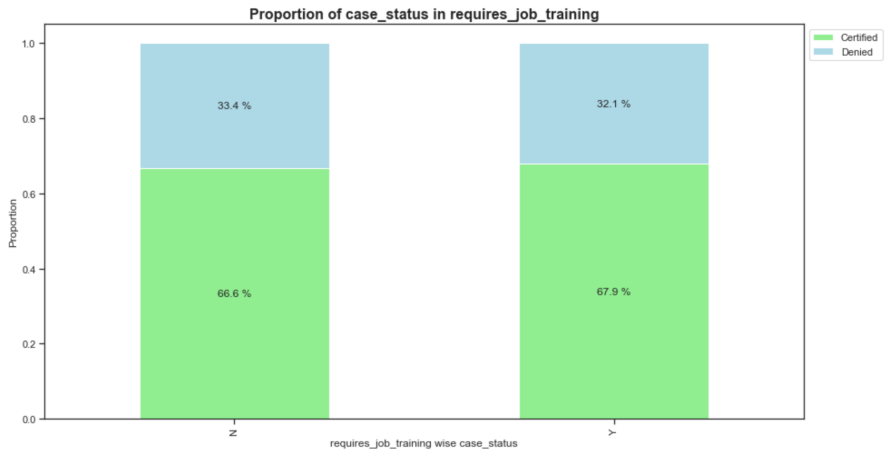
Case status by job experience



Notes:

1. Applicants with prior job experience shows higher approval rate ~75%
2. With no prior job experience the approval rate falls to ~56%

Case status by requires job training

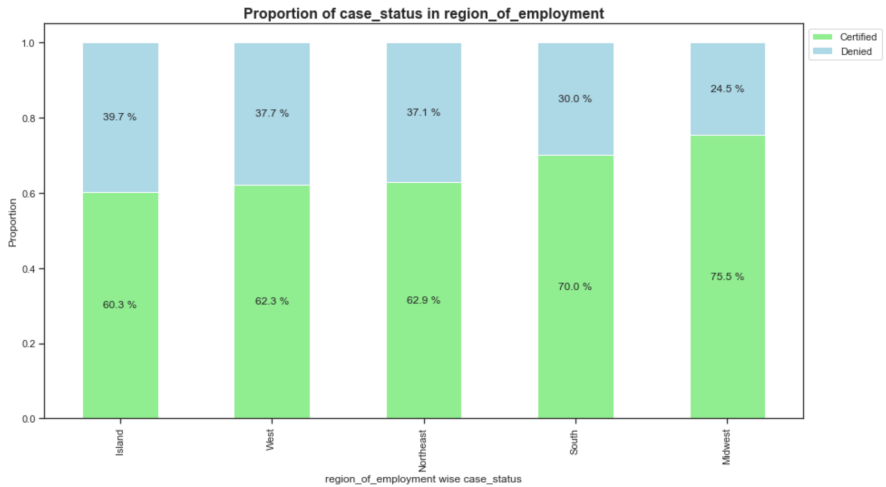


Notes:

1. No significance difference could be observed based on job training require or not
2. For both approval rate is close ~67-68%

Exploratory Data Analysis – Bivariate Analysis

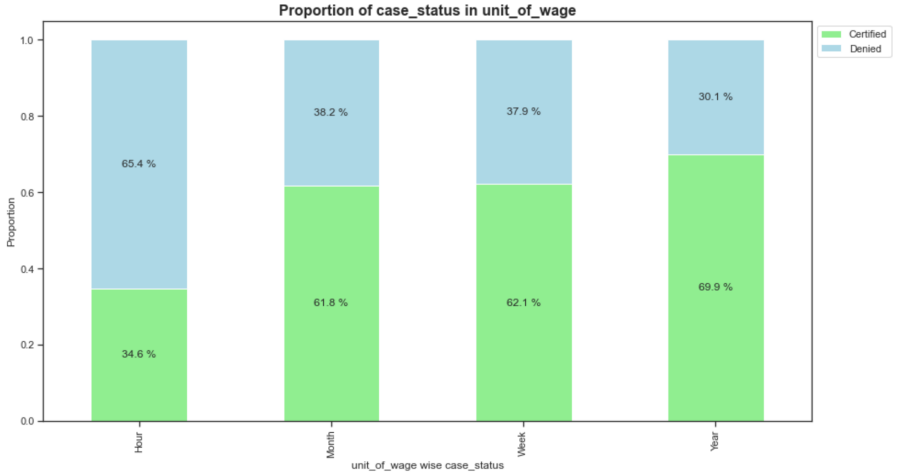
Case status by region of employment



Notes:

1. Case being approved or denied varies based on intended employment region
2. Intended region of employment Midwest has the higher approval rate ~76%
3. US Island shows the lowest approval rate ~60%

Case status by unit of wage

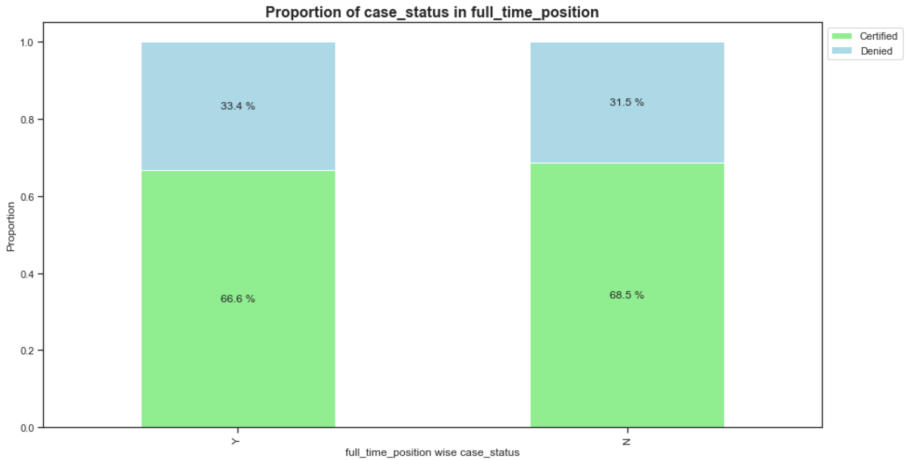


Notes:

1. Case status varies based on unit of prevailing wage
2. Yearly unit of wage shows the highest approval rate ~70% and hourly rate shows the lowest approval rate ~35%

Exploratory Data Analysis – Bivariate Analysis

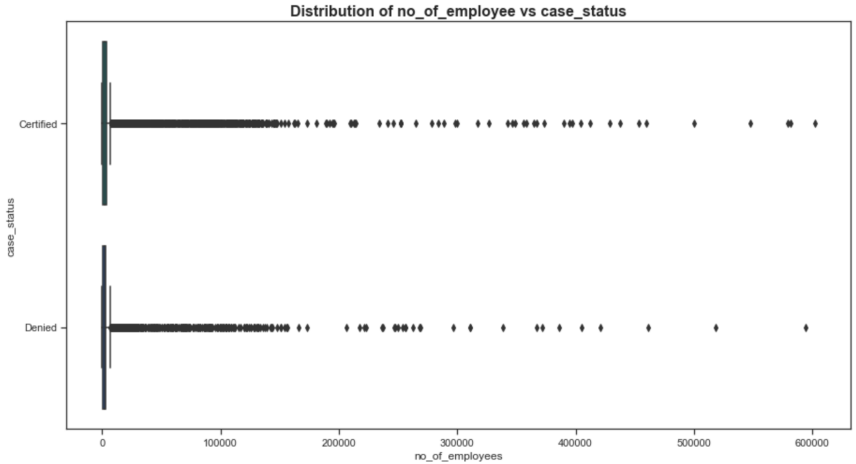
Case status by employment position fulltime or not



Notes:

1. No significant difference observed in approval rate based on applicant is applying for a fulltime position or not.
2. Approval rate is ~66-68% regardless positions which are full time or not.

Distribution of no of employee by case status

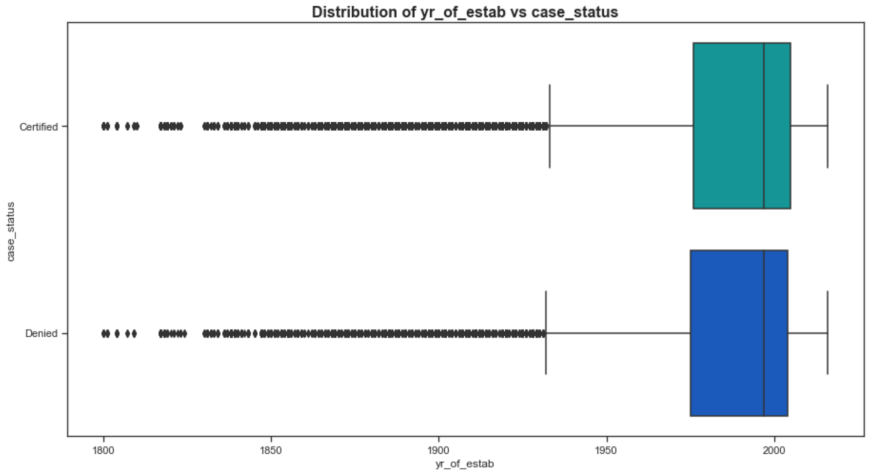


Notes:

1. No difference on average employee strength for each case status.
2. For both certified/denied status average employee strength is below ~5K

Exploratory Data Analysis – Bivariate Analysis

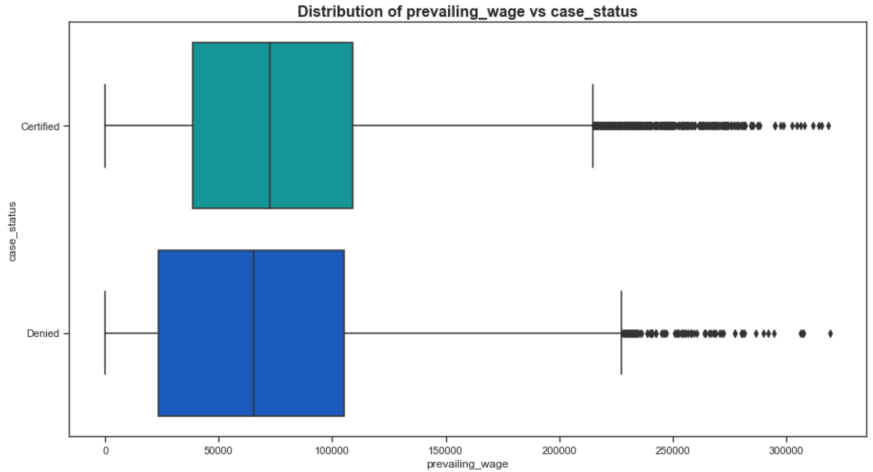
Distribution of year of establishment by case status



Notes:

1. No significant difference observed in year of establishment by case status.
2. For both case status we can observe the employers are established on and average ~1990 or after

Distribution of prevailing wage by case status



Notes:

1. Certified case status shows higher average prevailing wage than denied cases.

Model Performance Summary – Logistic Regression

Approach

- Explored Decision tree, Random Forest, Bagging, Adaboost, Gradient boosting, XGBoost, and Stacking Classifier for creating models for the solution.
- 30% of randomly chosen data are kept for testing model performance.
- Observation count for model training 17836, and testing 7644, with proportions of target variable similar to original dataset
- Model performance was checked against Recall, and F1 Score as for this problem we are trying to minimize our False negatives, which means trying to identify profiles which have potentials to get their visa certified, at the same time there should be a good balance between identifying both cases to minimize loss.

Model and Parameters

- Every model we explored for this solution are prepared with all independent variables except case id and year of establishment.
- Case id has been dropped because it's just an unique identifier for each case.
- Year of establishment showed majority as outliers and no significant difference between each case status, hence dropped.
- Except Stacking classifier each model was further tuned by finding best model parameters using GridSearchCV
- Below we can see the comparison of training and testing performance of each model

Performance of the Models

Training performance comparison:

	Decision Tree	Decision Tree Estimator	Random Forest Estimator	Random Forest Tuned	Bagging Classifier	Bagging Estimator Tuned	Adaboost Classifier	Adabosst Classifier Tuned	Gradient Boost Classifier	Gradient Boost Classifier Tuned	XGBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	1.0	0.737834	1.0	0.758354	0.982003	0.986208	0.737778	0.749776	0.756840	0.758746	0.825746	0.748710	0.760092
Recall	1.0	0.916226	1.0	0.918828	0.983296	0.998573	0.884244	0.892554	0.877193	0.877193	0.922773	0.917233	0.872240
Precision	1.0	0.747961	1.0	0.766044	0.989692	0.981113	0.761567	0.769615	0.784299	0.786305	0.834003	0.757609	0.790310
F1	1.0	0.823587	1.0	0.835509	0.986484	0.989766	0.818334	0.826538	0.828149	0.829266	0.876146	0.829815	0.829257

Testing performance comparison:

	Decision Tree	Decision Tree Estimator	Random Forest Estimator	Random Forest Tuned	Bagging Classifier	Bagging Estimator Tuned	Adaboost Classifier	Adabosst Classifier Tuned	Gradient Boost Classifier	Gradient Boost Classifier Tuned	XGBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	0.645997	0.728938	0.706698	0.737179	0.690476	0.724228	0.736264	0.741889	0.745814	0.748038	0.736002	0.737964	0.746991
Recall	0.734574	0.914006	0.817434	0.909109	0.767483	0.883839	0.884623	0.890695	0.872086	0.872674	0.861117	0.914789	0.866993
Precision	0.735150	0.740753	0.761080	0.750242	0.768687	0.748631	0.759886	0.762664	0.775340	0.777351	0.770552	0.748637	0.779088
F1	0.734862	0.818309	0.788251	0.822071	0.768085	0.810636	0.817524	0.821722	0.820872	0.822259	0.813321	0.823415	0.820693

Notes:

1. Most of the models has overfitted on the training data, tuning helped with overfitting except bagging classifier where even after hyperparameter tuning we have noticed the model overfitted.
2. Among all these models tuned XGBoost gave the highest f1 score in test data with ~82.3% and recall score ~91.4% which means close to 91% of applicants are correctly identified for whom the visa will be granted, and model performed well and in generalized manner, as the training and testing performances are stable and similar.
3. We concluded **XGBoost Tuned** is the best model and we can use this model to predict if VISA will be certified or denied based on profiles of applicant.

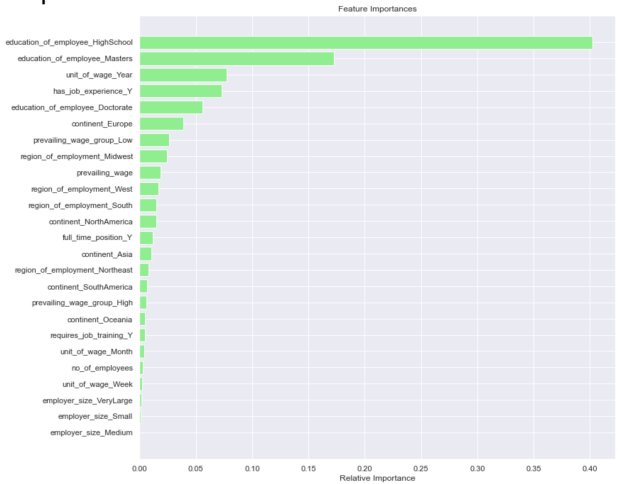
Model Performance Summary – XGBoost Tuned

Model and Parameters

- Initial model was prepared with default parameters and using XGBClassifier.
- Default XGBoost model showed sign of slight overfitting as we can see from the comparison result in the table on our right.
- This model is used for further tuning using GridSearchCV to reduce the overfit and increase performance
- Tuned XGBoost model gave generalized performance for both training and testing, and the model evaluation metric recall and f1 score are also pretty standard.
- Best parameters for tuned XGBoost model is -

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, eval_metric='logloss',
              gamma=0, gpu_id=-1, importance_type='gain',
              interaction_constraints='', learning_rate=0.300000012,
              max_delta_step=0, max_depth=6, min_child_weight=1, missing=nan,
              monotone_constraints='()', n_estimators=100, n_jobs=8,
              num_parallel_tree=1, random_state=1, reg_alpha=0, reg_lambda=1,
              scale_pos_weight=1, subsample=1, tree_method='exact',
              validate_parameters=1, verbosity=None)
```

Important Features



- Education of employee is the most important feature chosen by the model
- Unit of wage, prior job experience, continent of the applicant are some features which are shows significant importance for the model.
- We can observe there are some level of importance for almost all features fitted in the model

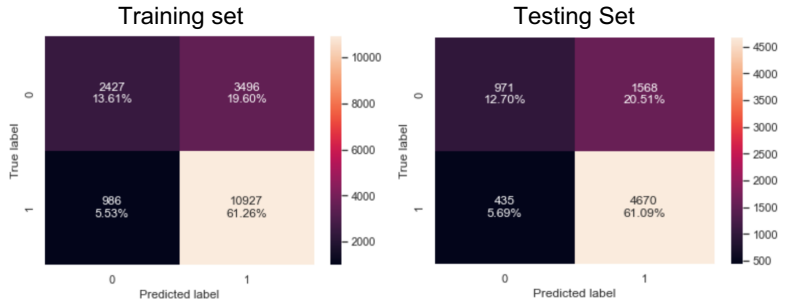
Performance of the Models

Models	Data	Recall	F1 Score
XGBoost – Default	Train	0.922773	0.876146
	Test	0.86117	0.813321
XGBoost - Tuned	Train	0.917233	0.829815
	Test	0.914789	0.823415

Notes:

- We can see model with best parameters identified by GridSearchCV able to generalize the model and eliminate overfitting
- Tuned XGBoost gave consistent performance between train and test with decent recall and f1 score

Confusion Matrix – Tuned XGBoost



Conclusion on Final Model

- Tuned XGBoost gave the highest f1 score in test data with ~82.3% and recall score ~91.4% which means close to 91% of applicants are correctly identified for whom the visa will be granted
- Hence, to address the business problem in hand we recommend to use the tuned XGBoost model

Business Insights and Recommendations

Insights

- Applicants having higher educations are more likely to have their case certified.
- Applicants having prior job experience and applying for a full-time position are more likely to have their case certified.
- Applicants having higher prevailing wage and yearly wage are more likely to have their case certified.
- We have observed if applicants from Europe has higher chances of their case getting certified.
- Applicants with intended work region as Midwest have higher chances of their case getting certified.
- XGBoost Tuned model gave robust and generalized performance. With recall score of ~91% and f1 score of ~82% model is predicting certified cases correctly 91% of the time, and both certified and denied cases approximately 82% of the time.

Recommendations

- Business would be able to use the finalized tuned model for getting profiles which are more likely to have their applications certified, this will help the approval process move faster as the number of applications goes higher.
- Model should be evaluated time to time and tuned as OFLC gathers more data over time.
- Capturing more attributes such as, job role, first time applicants/renewal, type of visa applications, for renewal change of visa types or not, for renewal if change of employer or not and analyzed further could lead to narrow down more precise profiles and if found significant add them to predictive model can make the model and predictive process more robust.

greatlearning
Power Ahead

Happy Learning !