EasyVisa - Ensemble Techniques

Michail Mersinias

04/25/2023

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

- The executive summary (insights) of the EasyVisa Ensemble Techniques presentation is as follows:
    - **Education Level:** An applicant applying for a job requiring a high school diploma will more than likely be denied. In contrast, applications for jobs requiring a Master's or a Doctorate degree are very likely to be approved.
    - **Prior Job Experience:** An applicant applying for a job without any previous job experience is more likely to be denied than an applicant for a job with experience.
    - **Prevailing Wage:** The higher the prevailing wage of the job an applicant is applying for, the more likely the application will be approved.
    - **Unit of Wage:** Applicants having a non-hourly (thus: weekly, monthly or yearly) unit of wage have higher chances of visa certification.
    - **Continent:** Applicants from Europe have higher chances of visa certification.

Executive Summary

- The executive summary (recommendations) of the EasyVisa Ensemble Techniques presentation is as follows:
    - First, sort applications by level of education and review the higher levels of education first.
    - Second, sort applications by previous job experience and review those with experience first.
    - Third, divide applications for jobs into those with an hourly wage and those with a non-hourly (weekly, monthly and yearly) wage, which constitute salaried jobs. Then, sort each group by the prevailing wage, and review applications for salaried jobs first from highest to lowest prevailing wage.
    - Fourth, do not take continent into account during classification in order to eliminate bias.
    - Finally, the model of choice can be a Tuned Random Forest Classifier, a Gradient Boosting Classifier, a Tuned XGBoost Classifier or a Stacking Classifier. They all perform similarly.

Business Problem Overview and Solution Approach

- Problem Definition:
  - The increasing number of work VISA applicants every year calls for a Machine Learning based solution that can help in shortlisting the candidates having higher chances of VISA approval.
  - The goal is to facilitate the process of visa approvals and recommend a suitable profile for the applicants for whom the visa should be certified or denied based on the drivers that significantly influence the case status.
- Solution Approach and Methodology:
  - For EDA, both univariate and multivariate analysis will be performed to find insights.
  - For Data Preprocessing, missing value imputation and feature engineering will be performed.
  - For Model Building, we will use the following classification models: Decision Tree, Bagging Classifier, Random Forest, AdaBoost, Gradient Boosting, XGBoost and the Stacking technique.

EDA Results: Data Overview

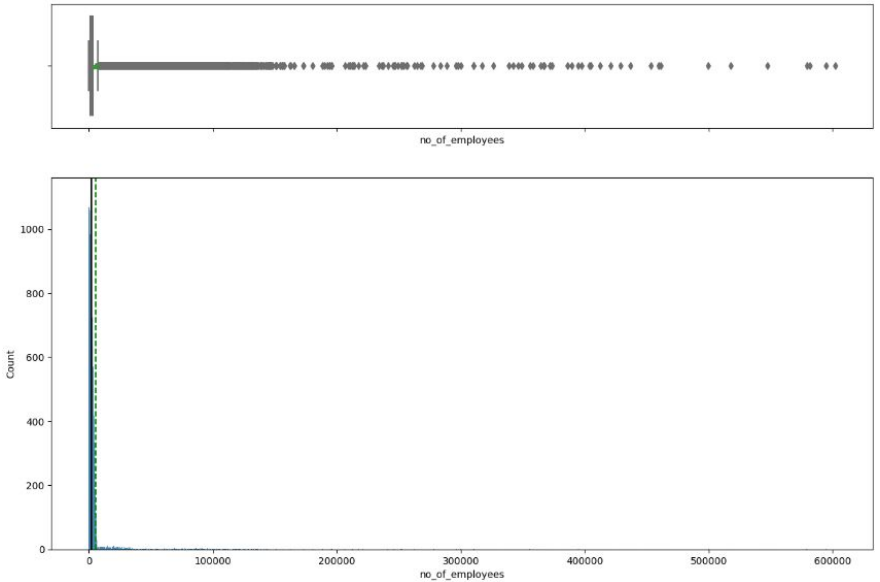
- The dataset is comprised of 25480 rows and 12 columns.
- The columns are as follows: case_id, continent, education_of_employee, has_job_experience, requires_job_training, no_of_employees, yr_of_estab, region_of_employment, prevailing_wage, unit_of_wage, full_time_position, case_status.
- All rows represent unique VISA cases (25480 unique cases).
- There are no duplicate values in the dataset.
- There are no missing values in the dataset.

EDA Results: Univariate Analysis

- In this section, we perform univariate analysis on the data.
  - For each attribute, we perform a descriptive statistical analysis where only that attribute is involved as a variable.
  - We also analyze the corresponding data and present the distribution of the attribute, with confidence intervals to signify variance and statistical significance.
  - Finally, we write the conclusion based on both quantitative and qualitative observations.

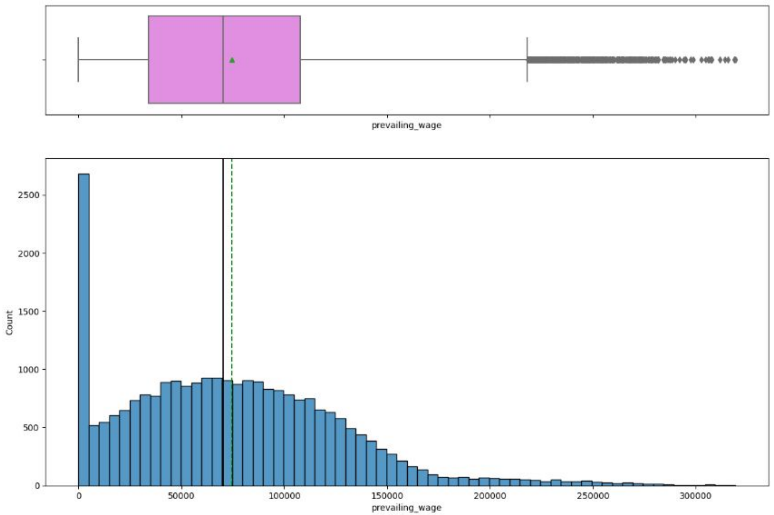
EDA Results: Univariate Analysis

- Analyzing the no_of_employees attribute, we report a mean value of 5667, with a standard deviation of 22877. Thus, we notice a significant variance and a significant number of outliers too. Two graphs, a histplot and a boxplot, are presented to depict this more clearly.



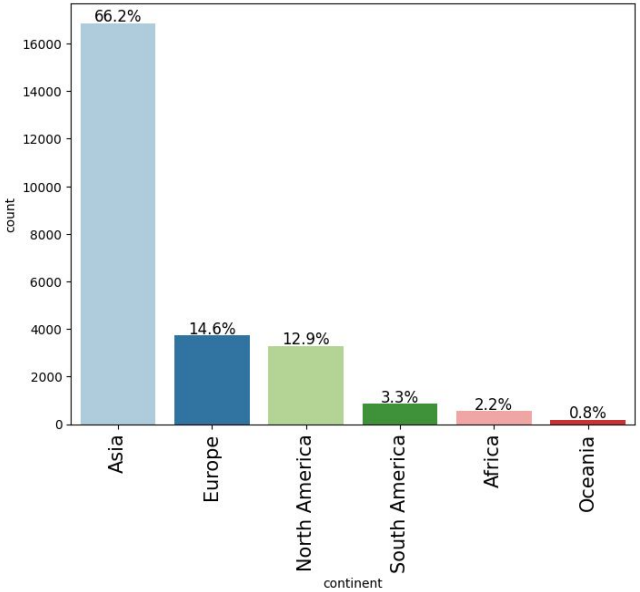
EDA Results: Univariate Analysis

- Analyzing the prevailing_wage attribute, we report a mean value of 74455, with a standard deviation of 52815. Thus, we notice a significant variance. We also notice a very high number of applicants close to zero, specifically 176 applicants have a prevailing_wage less than 100. Two graphs, a histplot and a boxplot, are presented to depict this more clearly.



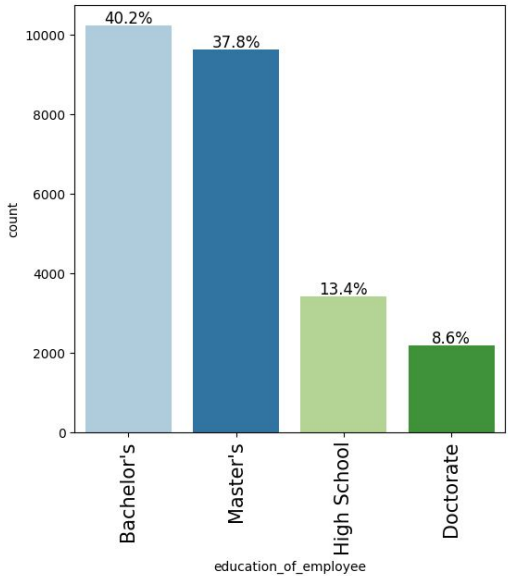
EDA Results: Univariate Analysis

- Analyzing the continent attribute, we report that most applications are from Asia (66.2%), followed by Europe (14.6%) and North America (12.9%). This makes sense, considering the population disparity of these continents. A countplot is presented to depict this more clearly.



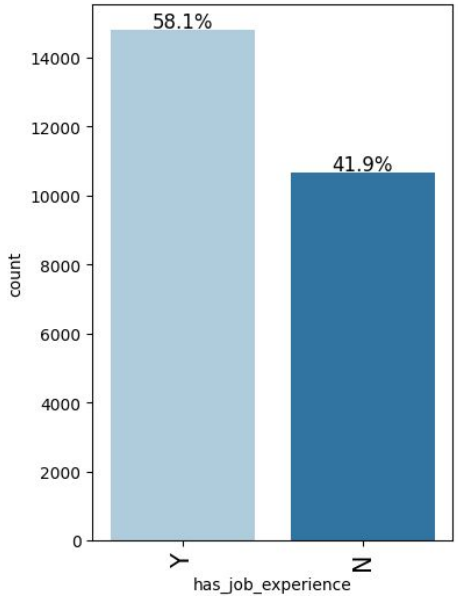
EDA Results: Univariate Analysis

- Analyzing the education_of_employee attribute, we report that most applicants have either a Bachelor (40.2%) or a Master (37.8%) degree. A countplot is presented to depict this more clearly.



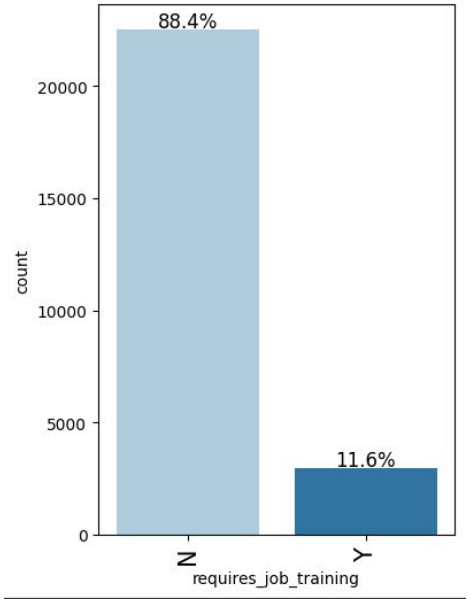
EDA Results: Univariate Analysis

- Analyzing the has_job_experience attribute, we report that 58.1% of applicants have job experience, while 41.9% of applicants don't. A countplot is presented to depict this more clearly.



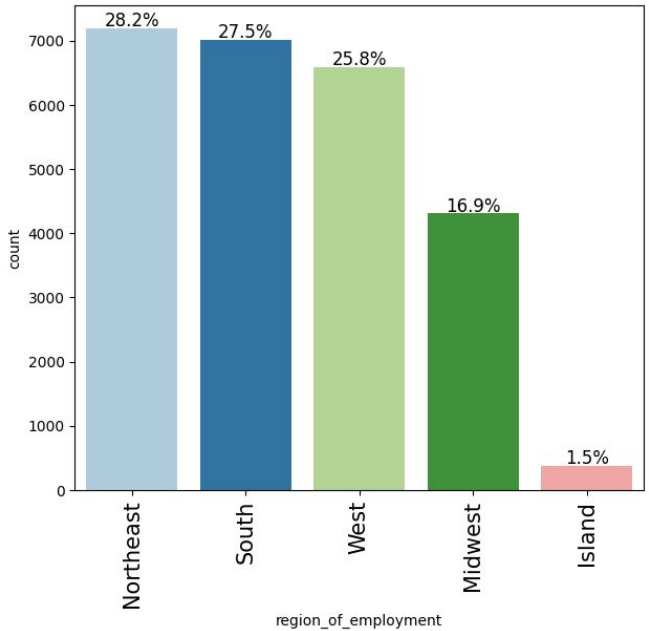
EDA Results: Univariate Analysis

- Analyzing the requires_job_training attribute, we report that most applicants (88.4%) do not require job training. A countplot is presented to depict this more clearly.



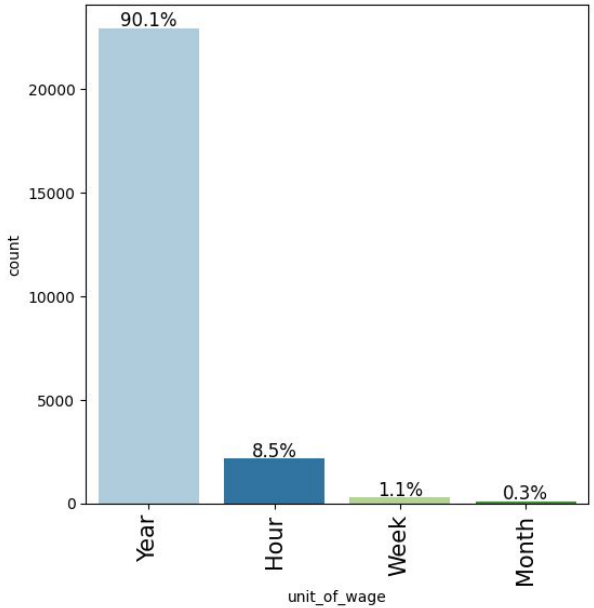
EDA Results: Univariate Analysis

- Analyzing the region_of_employment attribute, we report that the Northeast (28.2%), the South (27.5%) and the West (25.8%) are the top choices. A countplot is presented to depict this more clearly.



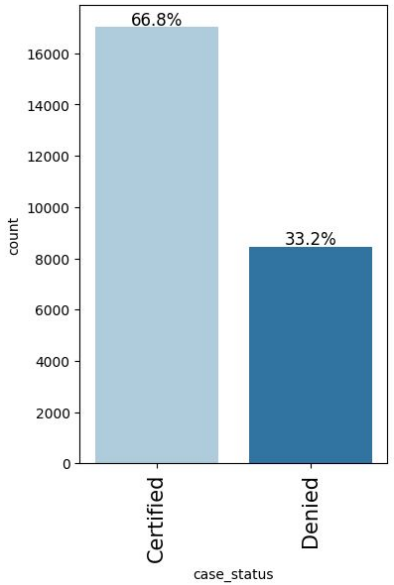
EDA Results: Univariate Analysis

- Analyzing the unit_of_wage attribute, we report that the majority of applications have a yearly unit of wage (90.1%), followed by hourly (8.5%). A countplot is presented to depict this more clearly.



EDA Results: Univariate Analysis

- Analyzing the case_status attribute, we that about two thirds of the applications (66.8%) were certified. A countplot is presented to depict this more clearly.

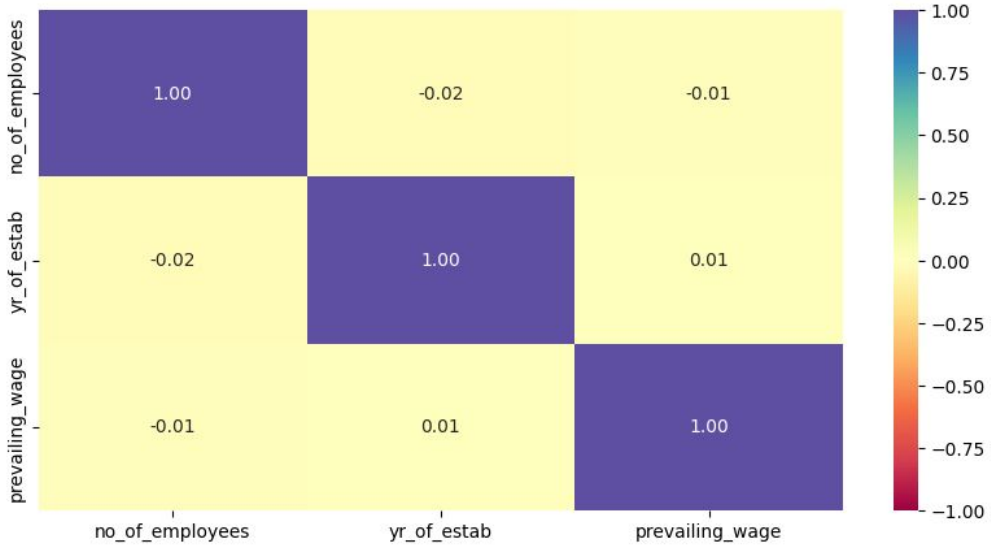


EDA Results: Multivariate Analysis

- In this section, we perform multivariate analysis on the data.
    - For selected attributes, we perform a descriptive statistical analysis where each attribute is jointly examined along with other attributes in order to evaluate their relationship and correlation.
    - We also analyze the corresponding data and present the joint distribution of the attributes, with confidence intervals to signify variance and statistical significance.
    - Finally, we write the conclusion based on both quantitative and qualitative observations.

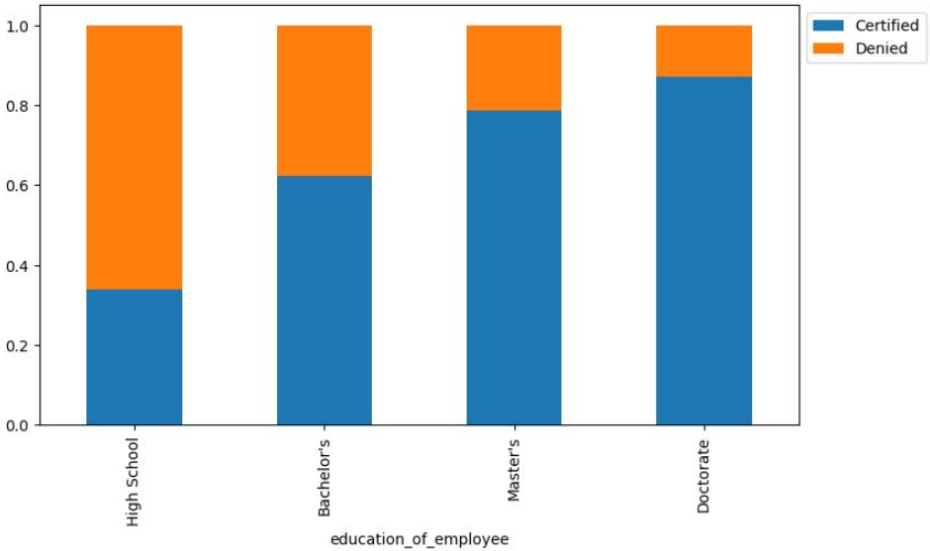
EDA Results: Multivariate Analysis

- Analyzing the correlation matrix between the variables no_of_employees, yr_of_estab and prevailing_wage, we cannot highlight any significant correlation between any of these variables.



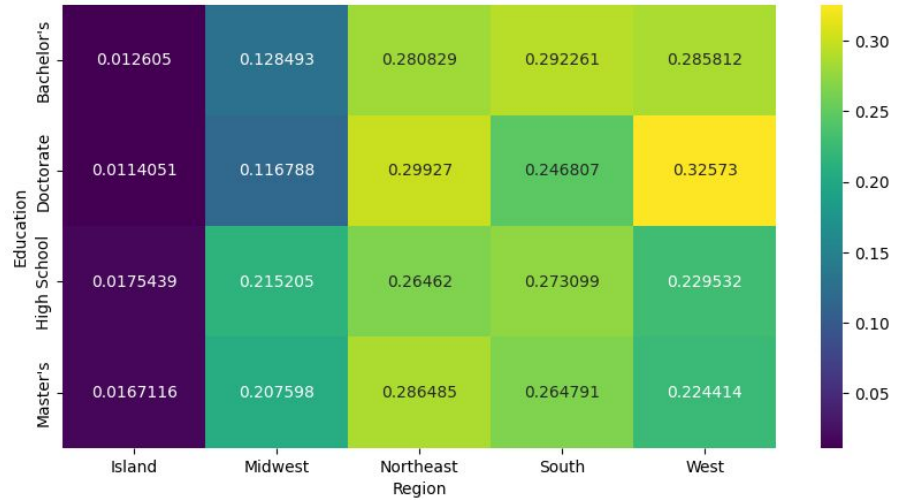
EDA Results: Multivariate Analysis

- Analyzing the relationship between the education_of_employee and case_status attributes, we observe that the higher the education of the employee, the higher chance for their visa applications to be accepted. Specifically, Master's (80%) and Doctorate (90%) achieve the highest probabilities.



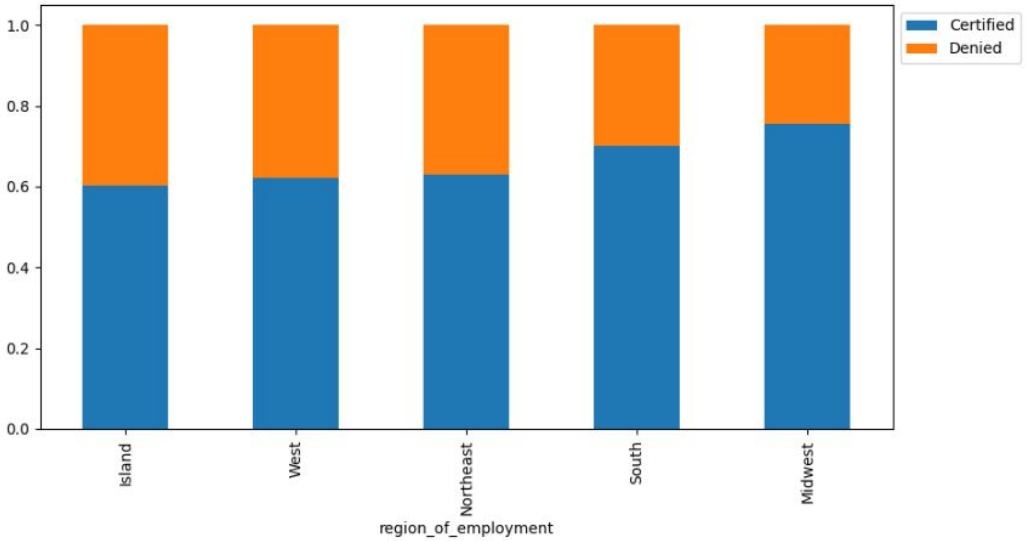
EDA Results: Multivariate Analysis

- Analyzing the relationship between the education_of_employee and region_of_employement attributes, we observe that those with a High School diploma have no strong regional preference. Those with a Bachelor's prefer the West, the South and the Northeast regions equally. Those with a Master's seem to prefer the South and the Northeast, while those with a Doctorate prefer the West and the Northeast.



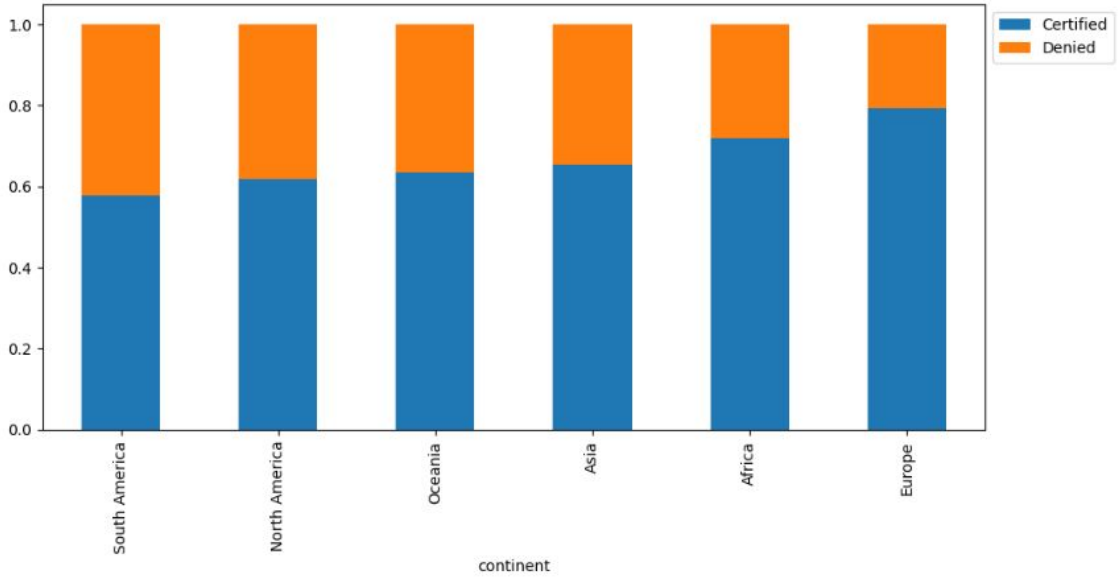
EDA Results: Multivariate Analysis

- Analyzing the relationship between the region_of_employment and case_status attributes, we observe that applications aimed towards the South and the Midwest regions have a slightly higher chance of acceptance (~70%), while the rest have a chance of acceptance around 60%.



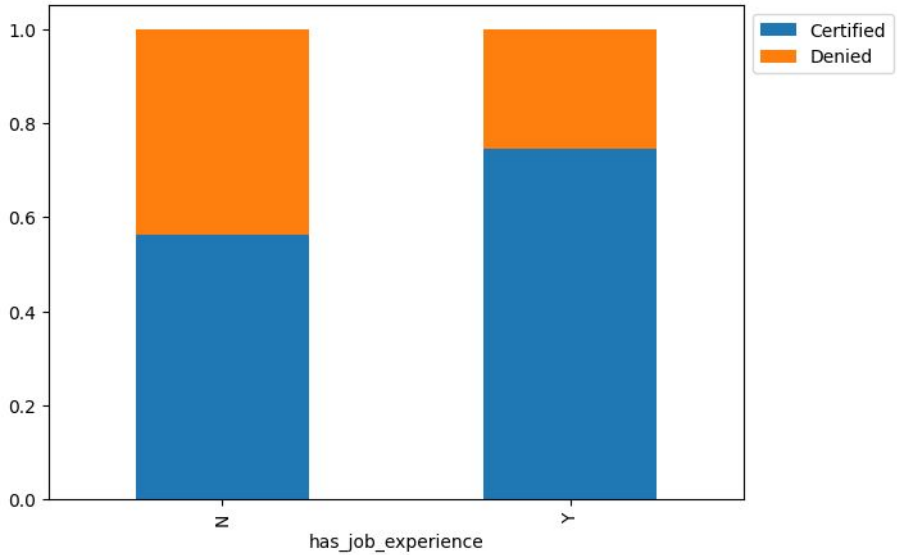
EDA Results: Multivariate Analysis

- Analyzing the relationship between the continent and case_status attributes, we observe that applicants from Europe (80%), Africa (75%) and Asia (70%) have the highest chance of being accepted.



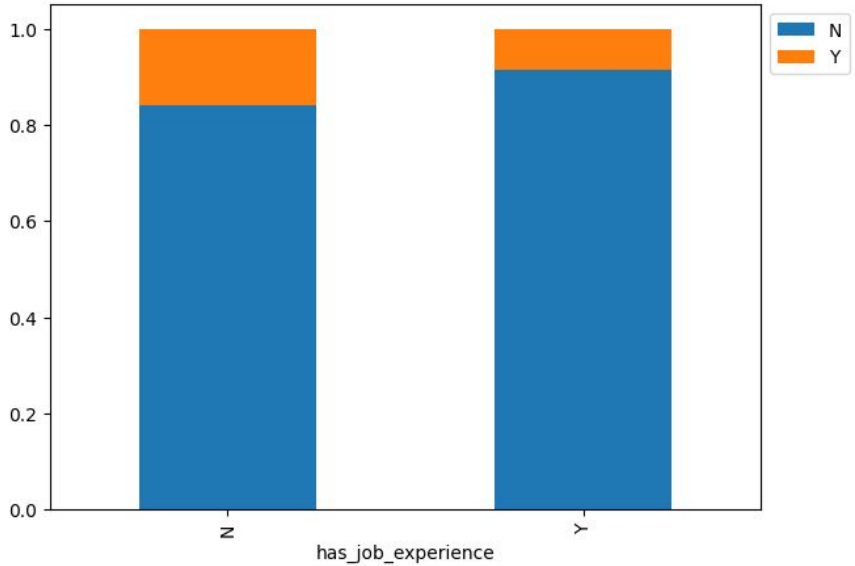
EDA Results: Multivariate Analysis

- Analyzing the relationship between the has_job_experience and case_status attributes, we observe that if the applicant has job experience, the chance for their application to be accepted is approximately 20% higher.



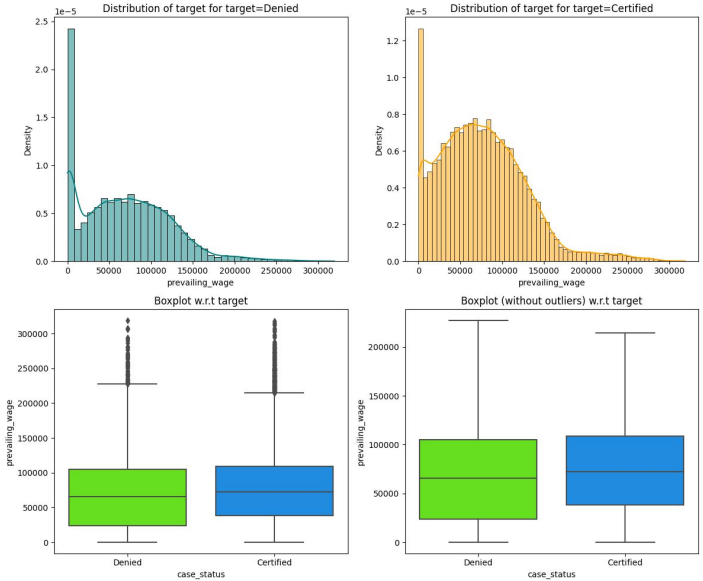
EDA Results: Multivariate Analysis

- Analyzing the relationship between the has_job_experience and requires_job_training attributes, we observe that the applicants who have prior job experience are approximately 10% more likely not to require extra job training.



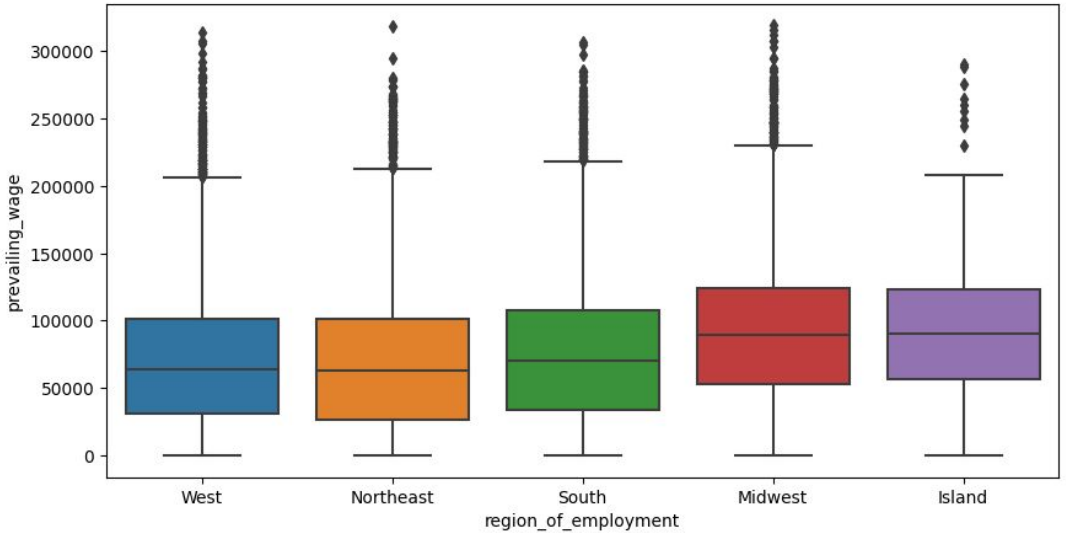
EDA Results: Multivariate Analysis

- Analyzing the relationship between the prevailing_wage and case_status attributes, we observe that the accepted applicants have a significantly higher prevailing wage. Furthermore, a significant number of denied applications have a prevailing wage close to zero.



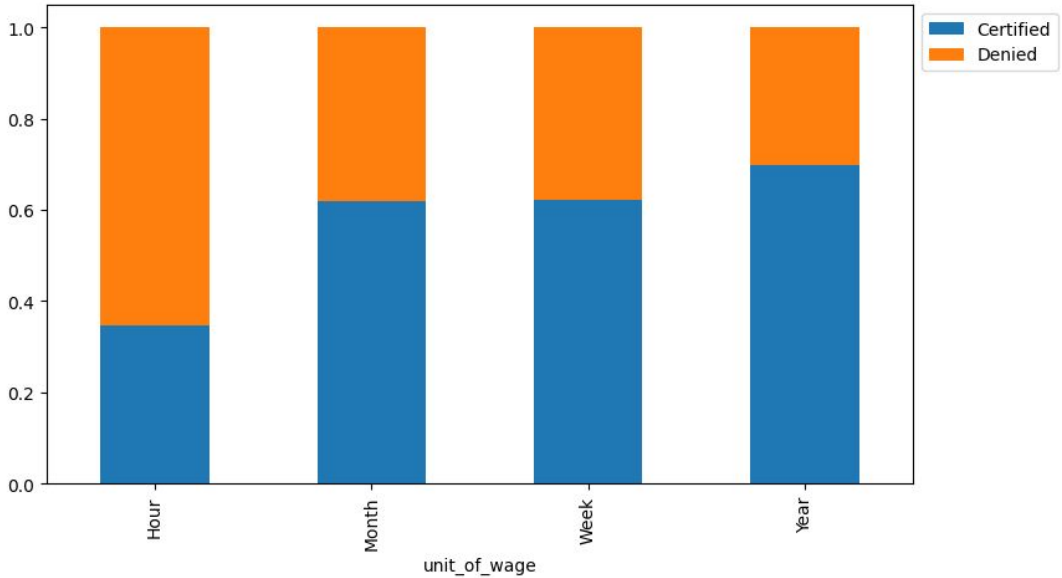
EDA Results: Multivariate Analysis

- Analyzing the relationship between the region_of_employment and prevailing_wage attributes, we observe no significant variations of prevailing wages across regions of employment.



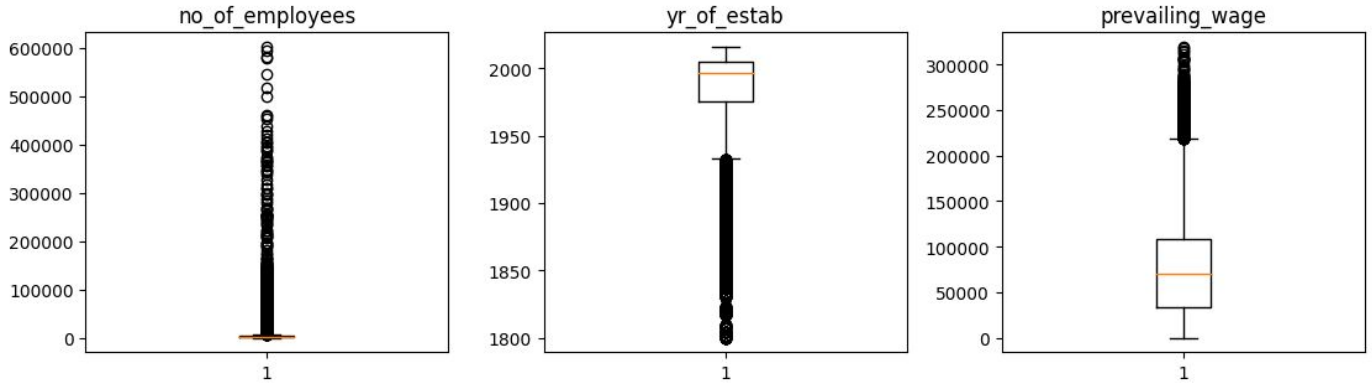
EDA Results: Multivariate Analysis

- Analyzing the relationship between the unit_of_wage and case_status attributes, we observe that a the applicants whose unit of wage is not hourly have a significantly higher chance to have their applications accepted.



Data Preprocessing

- Outlier Check: There are numerous outliers in this most columns of this dataset, however no action will be taken as they are accurate values that contain important information about each feature.



- Preparing Data for Modeling: We use case_status as our target variable which we seek to predict accurately, and all the other columns as features to train our model.
- Train/Test Dataset Split: We split the data in 70:30 ratio for train to test data.

Model Performance Summary

- Modeling Setup:
  - We use the following classification models: Decision Tree, Bagging Classifier, Random Forest, AdaBoost, Gradient Boosting, XGBoost and the Stacking technique.
  - For each one of these models, we use GridSearchCV with cv = 5, in order to find the most suitable model parameters based on the training data.
  - Finally, we define functions for the following metrics of performance: Accuracy, Precision, Recall and F1 Score.

Model Performance Summary

- The results on the training data are the following:

	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	1.0	0.712548	0.985198	0.996187	1.0	0.769119	0.738226	0.718995	0.758802	0.764017	0.838753	0.765474	0.770688
Recall	1.0	0.931923	0.985982	0.999916	1.0	0.918660	0.887182	0.781247	0.883740	0.882649	0.931419	0.881642	0.892554
Precision	1.0	0.720067	0.991810	0.994407	1.0	0.776556	0.760688	0.794587	0.783042	0.789059	0.843482	0.791127	0.790969
F1	1.0	0.812411	0.988887	0.997154	1.0	0.841652	0.819080	0.787861	0.830349	0.833234	0.885272	0.833935	0.838697

- Therefore, for the training data, a Decision Tree Classifier and a Random Forest Classifier offer the highest performance, but is is quite likely that this high performance is a result of overfitting.

Model Performance Summary

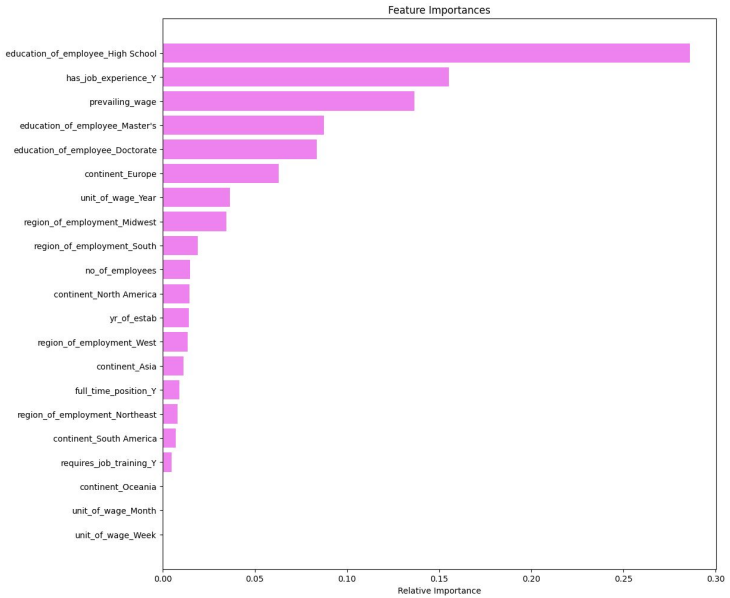
- The results on the test data are the following:

	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	0.664835	0.706567	0.691523	0.724228	0.727368	0.738095	0.734301	0.716510	0.744767	0.743459	0.733255	0.745160	0.745160
Recall	0.742801	0.930852	0.764153	0.895397	0.847209	0.898923	0.885015	0.781391	0.876004	0.871303	0.860725	0.869540	0.879726
Precision	0.752232	0.715447	0.771711	0.743857	0.768343	0.755391	0.757799	0.791468	0.772366	0.773296	0.767913	0.775913	0.770987
F1	0.747487	0.809058	0.767913	0.812622	0.805851	0.820930	0.816481	0.786397	0.820927	0.819379	0.811675	0.820063	0.821775

- Therefore, for the test data, a Tuned Random Forest Classifier, a Gradient Boosting Classifier, a Tuned XGBoost Classifier and a Stacking Classifier offer the highest performance with an F-1 score of approximately 82%.

Model Performance Summary

- The feature importance of some the top-performing models (Gradient Boosting Classifier, XGBoost Classifier) is approximately the same and is depicted in the following figure:



- Important features (Importance > 0.05): education_of_employee, has_job_experience, prevailing_wage, continent, unit_of_wage.

Executive Summary

- The executive summary (insights) of the EasyVisa Ensemble Techniques presentation is as follows:
    - **Education Level:** An applicant applying for a job requiring a high school diploma will more than likely be denied. In contrast, applications for jobs requiring a Master's or a Doctorate degree are very likely to be approved.
    - **Prior Job Experience:** An applicant applying for a job without any previous job experience is more likely to be denied than an applicant for a job with experience.
    - **Prevailing Wage:** The higher the prevailing wage of the job an applicant is applying for, the more likely the application will be approved.
    - **Unit of Wage:** Applicants having a non-hourly (thus: weekly, monthly or yearly) unit of wage have higher chances of visa certification.
    - **Continent:** Applicants from Europe have higher chances of visa certification.

Executive Summary

- The executive summary (recommendations) of the EasyVisa Ensemble Techniques presentation is as follows:
    - First, sort applications by level of education and review the higher levels of education first.
    - Second, sort applications by previous job experience and review those with experience first.
    - Third, divide applications for jobs into those with an hourly wage and those with a non-hourly (weekly, monthly and yearly) wage, sort each group by the prevailing wage, then review applications for salaried jobs first from highest to lowest wage.
    - Fourth, do not take continent into account during classification in order to eliminate bias.
    - Finally, the model of choice can be a Tuned Random Forest Classifier, a Gradient Boosting Classifier, a Tuned XGBoost Classifier or a Stacking Classifier. They all perform similarly.

Thank you for your time!

Michail Mersinias

04/25/2023