← Go Back to Ensemble Techniques

Course Content

EasyVisa - Problem Statement

Submission type	:	File Upload
Due Date	:	Jan 20, 1:30 AM CET
Total Marks	:	60
Available from	:	Jan 04, 3:30 PM

Description

Context

Business communities in the United States are facing high demand for human resources, but one of the constant challenges is identifying and attracting the right talent, which is perhaps the most important element in remaining competitive. Companies in the United States look for hard-working, talented, and qualified individuals both locally as well as abroad.

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. The immigration programs are administered by the Office of Foreign Labor Certification (OFLC).

OFLC processes job certification applications for employers seeking to bring foreign workers into the United States and grants certifications in those cases where employers can demonstrate that there are not sufficient US workers available to perform the work at wages that meet or exceed the wage paid for the occupation in the area of intended employment.

Objective

In 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. OFLC has hired the firm EasyVisa for data-driven solutions. You as a data scientist at EasyVisa have to analyze the data provided and, with the help of a classification model:

1. Facilitate the process of visa approvals.
2. 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.

Data Description

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

case_id: ID of each visa application
continent: Information of continent the employee
education_of_employee: Information of education of the employee
has_job_experience: Does the employee has any job experience? Y= Yes; N = No
requires_job_training: Does the employee require any job training? Y = Yes; N = No
no_of_employees: Number of employees in the employer's company
yr_of_estab: Year in which the employer's company was established
region_of_employment: Information of foreign worker's intended region of employment in the US.
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.
unit_of_wage: Unit of prevailing wage. Values include Hourly, Weekly, Monthly, and Yearly.
full_time_position: Is the position of work full-time? Y = Full-Time Position; N = Part-Time Position
case_status: Flag indicating if the Visa was certified or denied

Note:

Please note XGBoost can take a significantly longer time to run, so if you have time complexity issues then you can avoid building and tuning XGBoost. No marks will be deducted if the XGBoost model is not attempted.

Submission Guidelines

1. There are two ways to work on this project:

i. Full-code way: The full code way is to write the solution code from scratch and only submit a final Jupyter notebook with all the insights and observations.

ii. Low-code way: The low-code way is to use an existing solution notebook template to build the solution and then submit a business presentation with insights and recommendations.

The primary purpose of providing these two options is to allow learners to opt for the approach that aligns with their individual learning aspirations and outcomes. The below table elaborates on these two options.

Submission type	Who should choose	What is the same across the two	What is different across the two	Final submission file [IMP]	Submission Format
Full-code	Learners who aspire to be in hands-on coding roles in the future focussed on building solution codes from scratch	Perform exploratory data analysis to identify insights and recommendations for the problem	Focus on code writing: 10-20% grading on the quality of the final code submitted	Solution notebook from the full-code template submitted in .html format	.html
Low-code	Learners who aspire to be in managerial roles in the future-focussed on solution review, interpretation, recommendations, and communicating with business		Focus on business presentation: 10-20% grading on the quality of the final business presentation submitted	Business presentation in .pdf format with problem definition, insights, and recommendations	.pdf

Please follow the below steps to complete the assessment. **Kindly note that if you submit a presentation along with the notebook, ONLY the presentation will be evaluated.** Please make sure that all the sections mentioned in the rubric have been covered in your submission.

i. Full-code version

Download the full-code version of the learner notebook.
Follow the instructions provided in the notebook to complete the project.
Clearly write down insights and recommendations for the business problems in the comments.
Submit only the solution notebook prepared from the learner notebook [format: .html]

ii. Low-code version

Download the low-code version of the learner notebook.

Follow the instructions provided in the notebook to complete the project.

Prepare a business presentation with insights and recommendations to the business problem.

Submit only the presentation [format: .pdf]

2. Any assignment found copied/plagiarized with other submissions will not be graded and awarded zero marks.

3. Please ensure timely submission as any submission post-deadline will not be accepted for evaluation.

4. Submission will not be evaluated if

it is submitted post-deadline, or,

more than 1 file is submitted.

Best Practices for Full-code submissions

The final notebook should be well-documented, with inline comments explaining the functionality of code and markdown cells containing comments on the observations and insights.
The notebook should be run from start to finish in a sequential manner before submission.

It is important to remove all warnings and errors before submission.
The notebook should be submitted as an HTML file (.html) and NOT as a notebook file (.ipynb).
Please refer to the FAQ page for common project-related queries.

Best Practices for Low-code submissions

The presentation should be made keeping in mind that the audience will be the Data Science lead of a company.
The key points in the presentation should be the following:
- Business Overview of the problem and solution approach
- Key findings and insights which can drive business decisions
- Business recommendations
- Focus on explaining the key takeaways in an easy-to-understand manner.
- The inclusion of the potential benefits of implementing the solution will give you the edge.

Copying and pasting from the notebook is not a good idea, and it is better to avoid showing codes unless they are the focal point of your presentation.
The presentation should be submitted as a PDF file (.pdf) and NOT as a .pptx file.
Please refer to the FAQ page for common project-related queries.

Happy Learning!

Scoring guide (Rubric) - EasyVisa Rubric

Criteria	Points
Exploratory Data Analysis - Problem definition - Univariate analysis - Bivariate analysis - Use appropriate visualizations to identify the patterns and insights - Key meaningful observations on individual variables and the relationship between variables	11
Data Pre-processing	4

Criteria	Points
- Duplicate value check - Missing value treatment (if needed) - Outlier Detection(treat, if needed- why or why not), - Feature Engineering (if needed) - Data preparation for modeling	
Model Building - Bagging - Build Decision Tree, Bagging classifier, and Random Forest - Comment on model performance	7
Model Improvement - Bagging - Comment on the model performance after tuning the Decision Tree, Bagging, and Random Forest classifier to improve the model performance.	10
Model Building - Boosting - Build Adaboost and GradientBoost - Comment on model performance * Please note building XGBoost model is optional	5
Model Improvement - Boosting - Comment on the model performance after tuning the AdaBoost, and Gradient Boosting classifier on the appropriate metric to improve the model performance - Build Stacking Classifier * Please note XGBoost can take a significantly longer time to run, so if you have time complexity issues then you can avoid tuning XGBoost.	9
Actionable Insights & Recommendations - Compare model performance on various metrics. - Conclude with the key takeaways - What would your advice be to grow the business?	6
Presentation / Notebook - Overall quality - Structure and flow - Crispness - Visual appeal - Conclusion and Business Recommendations OR - Structure and flow - Well	8

Criteria	Points
commented code - Conclusion and Business Recommendations	
Points	60

AI Mentor
Chat with AI Mentor to get instant answers to your queries related to this assignment.

Submit Assignment

Upload file

Add comments

Submit Assignment