BUSINESS REPORT ON

EASYVISA PROJECT

TABLE OF CONTENTS

TITLE **PAGE**

NO

LIST OF FIGURES

TITLE **PAGE NO**

1. OBJECTIVE

The Easy Visa project is to support the U.S. Office of Foreign Labor Certification (OFLC) in efficiently processing visa applications by developing a **machine learning model** that can predict the likelihood of a visa application being certified or denied..

2.BUSINESS CONTEXT

The Easy Visa project centres around the challenges that U.S. businesses face in addressing workforce shortages while navigating the complexities of immigration law and labour certification. The demand for skilled labour in the United States is high, and businesses often seek qualified talent from both local and international sources to remain competitive in the global market. However, hiring foreign workers involves complying with the **Immigration and Nationality Act (INA)** and regulations enforced by the **Office of Foreign Labor Certification (OFLC)**, which seeks to protect U.S. workers by ensuring fair wages and working conditions.

To address this challenge, Easy Visa, a data science consulting firm, has been tasked with designing a **data-driven, machine learning solution** that can assist the OFLC by predicting the likelihood of visa certification or denial based on key applicant and employer characteristics. The solution aims to reduce the time required to process each application by quickly identifying high- and low-probability cases, allowing OFLC staff to focus their efforts where they're most needed. Additionally, the solution will provide insights into the factors that impact visa approval outcomes, helping employers better understand how they can structure applications to meet OFLC requirements.

3.DATA DESCRIPTION

The data contains the different factors to analyze for the content. The detailed data dictionary is given below.

Data Dictionary

- 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

4.EXPLORATORY DATA ANALYSIS (EDA)

The distribution and relationships of variables within the Easy Visa dataset. This analysis helps uncover insights about visa approval patterns that will inform the model-building process.

4.1 UNIVARIATE ANALYSIS

- **no_of_employees**
  - *Distribution*: The no_of_employees variable is right-skewed, indicating that most companies are relatively small in size, with a few significantly larger companies.
  - *Outliers*: There are outliers with very high employee counts, indicating the presence of large organizations, which could influence hiring practices and visa approvals.

- **yr_of_estab**
  - *Distribution*: The distribution of yr_of_estab appears fairly uniform, with companies established over a wide range of years.
  - *Outliers*: The presence of outliers suggests that some companies are either very new or very old, possibly affecting their hiring power or reputation in the visa process.
- **region_of_employment**
  - *Distribution*: There is a noticeable geographical variation, with certain regions showing higher frequencies of employment cases. This might reflect regional labor demands and immigration patterns.
- **prevailing_wage**
  - *Distribution*: This feature is right-skewed, with most prevailing wages falling on the lower end and a few high outliers.
  - *Outliers*: Very high prevailing wages may be associated with specialized positions or higher-cost regions, which could affect visa approval likelihood.
- **unit_of_wage**
  - *Distribution*: The majority of wages are measured in "Year" units, providing insights into common wage reporting standards and allowing consistent comparisons.
- **full_time_position**
  - *Distribution*: A higher proportion of applications are for full-time positions, aligning with typical visa applications, where full-time work is a standard requirement.
- **education_of_employee**
  - *Distribution*: Certain education levels are more common among applicants, indicating that many visa applications may come from fields with specific educational prerequisites.
- **has_job_experience**

- *Distribution*: Most applicants have prior job experience, suggesting that work experience could be a favorable factor in the visa application process.

➢ **requires_job_training**

- *Distribution*: A smaller subset of cases require job training, indicating that most positions do not mandate job training for applicants.

➢ **case_status**

- *Distribution*: The target variable case_status shows an imbalance, with more cases being certified than denied. Addressing this imbalance will be crucial during model development to avoid biased predictions.

4.2 BIVARIATE ANALYSIS

Key Relationships between Case Status and Other Variables:

Prevailing Wage:

Insight: Certified cases tend to have a slightly higher median prevailing wage compared to denied cases.

Interpretation: This could indicate that higher wages are associated with a higher likelihood of certification, possibly due to the perception of greater economic value or the role's higher skill level.

Number of Employees:

Insight: Certified cases appear to be linked with employers that have a slightly larger workforce.

Interpretation: Larger companies might have a higher rate of certification, possibly due to their resources, reputation, or previous successful applications, making them favorable candidates.

Year of Establishment:

Insight: Certified cases tend to be associated with companies that were established earlier compared to those in denied cases.

Interpretation: This suggests that well-established companies might have a higher success rate, potentially due to established credibility and stable business operations.

Full-Time Position:

Insight: Full-time positions are more frequently associated with certified cases.

Interpretation: Applications for full-time roles may be more likely to meet visa certification criteria, perhaps due to the long-term economic impact of such positions.

Has Job Experience:

Insight: Cases for candidates with prior job experience are more likely to be certified.

Interpretation: Job experience might be viewed positively in the application process, signaling a candidate's readiness and lower need for training, making them favorable for certification.

Requires Job Training:

Insight: Cases requiring job training are less likely to be certified compared to those that don’t require training.

Interpretation: Job training requirements could indicate a lack of preparedness or specific skills, which might reduce the likelihood of certification as it may be seen as a risk.

Education of Employee:

Insight: Certain education levels are associated with higher certification rates.

Interpretation: This could highlight preferred education levels for certification, possibly reflecting job requirements or perceived skill levels.

Region of Employment:

Insight: Certification rates vary by employment region.

Interpretation: Regional differences in labor demand or visa policies could influence certification rates, highlighting regions that may need more or fewer foreign workers.

Unit of Wage:

Insight: Certification rates may differ based on the unit of wage (e.g., hourly, yearly).

Interpretation: Certain wage units might be more favorable, perhaps reflecting job duration expectations or wage stability.

Correlation Matrix Insights:

The correlation matrix helps identify relationships among numerical variables, offering insights into potential interactions:

- For example, if prevailing_wage and no_of_employees show a positive correlation, it suggests that larger companies tend to offer higher wages.

- This can further explain why larger companies might have higher certification rates, as they tend to offer wages that align with certification standards.

4.3 CORRELATION MATRIX

- Prevailing Wage and Full-Time Position:

  Correlation: There is a negative correlation (-0.20) between prevailing_wage and full_time_position.

  Insight: This may suggest that full-time positions sometimes offer lower wages compared to part-time or specialized roles. This could be due to industry-specific factors, such as part-time roles in specialized fields offering higher hourly wages.

- Has Job Experience and Case Status:

  Correlation: There is a positive correlation (0.19) between has_job_experience and case_status.

  Insight: Job experience has a positive association with the likelihood of certification. This reinforces the idea that prior experience may make applicants more favorable candidates, potentially due to reduced training needs or proven skills.

- Full-Time Position and Prevailing Wage:

  Correlation: There is a slight negative correlation (-0.20) between full_time_position and prevailing_wage.

  Insight: This suggests that positions marked as full-time may tend to offer lower wages than part-time or specialized roles, where the wage might be higher due to unique skill requirements.

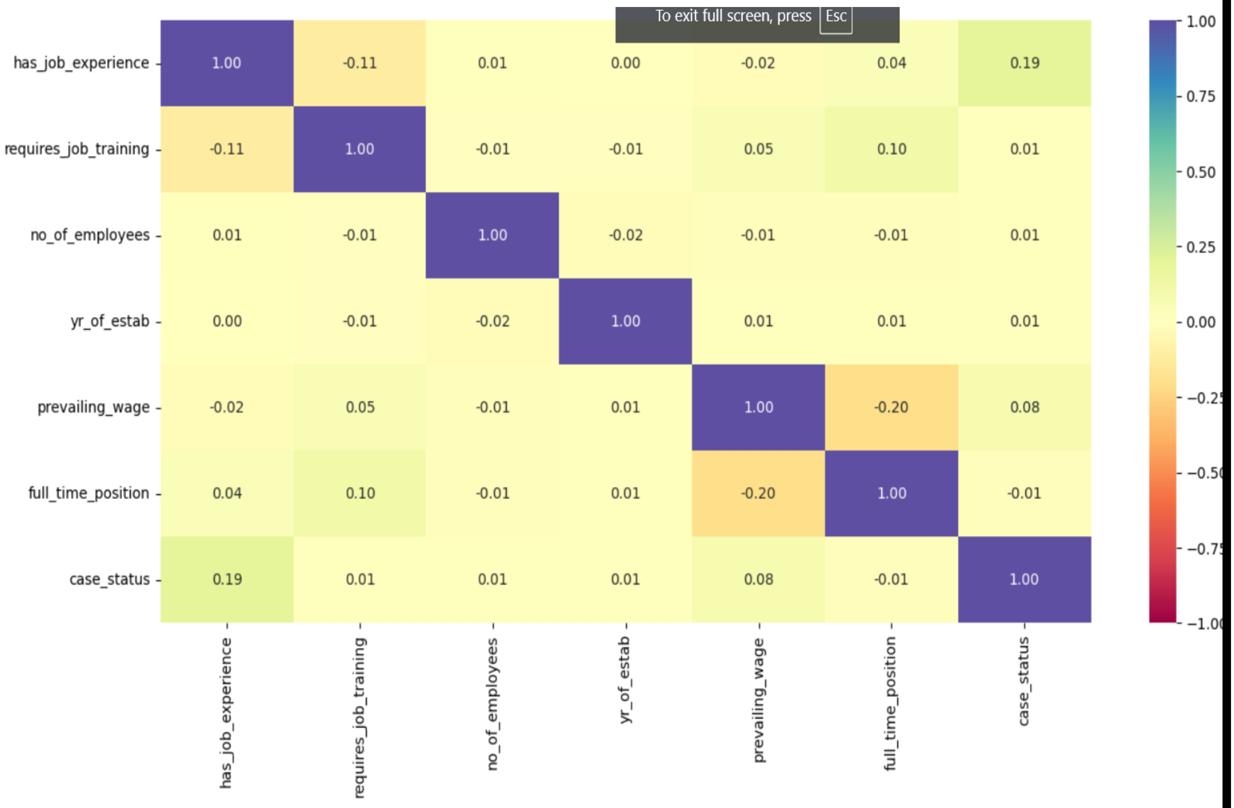
- No Significant Correlations:

  Observation: Most other variables show low or near-zero correlations with each other, indicating limited or weak linear relationships.

Insight: The low correlations suggest that many variables are independent, which can be beneficial for model training as it reduces multicollinearity concerns. However, this also suggests that individual variables may not have strong predictive power on their own, reinforcing the importance of using more complex algorithms that can capture non-linear relationships.

- weak correlations in the matrix suggest that many variables do not have strong linear relationships, but this doesn't rule out potential non-linear interactions. Therefore, using algorithms capable of capturing non-linear relationships (e.g., decision trees, gradient boosting) may be beneficial in the modeling stage.
- The slight positive correlation between has_job_experience and case_status highlights job experience as a potential driver for certification, while the negative correlation between full_time_position and prevailing_wage points to possible wage disparities based on employment type.

FIG 1 CORRELATION MATRIX



5. DATA PREPROCESSING

The data preparation steps outlined in this section are crucial for ensuring that the dataset is in a suitable format for modeling. By carefully separating features from the target variable, encoding categorical data, splitting the dataset into training and testing sets, and analyzing class distribution, we set a solid foundation for the subsequent modeling phase. This preparation enhances the model's ability to learn and generalize from the data, ultimately leading to better performance and insights.

6. OUTLIER ANALYSIS INSIGHTS

- Number of Employees:

  Outliers Detected: 1,556

  Insight: A significant number of companies have an unusually high or low number of employees compared to the rest of the dataset. These outliers could represent very large corporations or small niche firms. Large firms may have a greater capacity to employ foreign workers, which might influence visa application success rates.

- Year of Establishment:

  Outliers Detected: 3,260

  Insight: The presence of many outliers in the yr_of_estab variable suggests that there are numerous very old or very new companies in the dataset. This could imply that older, more established companies might have more credibility in visa applications, while very new companies could face additional scrutiny.

- Prevailing Wage:

  Outliers Detected: 427

  Insight: High or low prevailing wage outliers might represent specialized, high-paying roles or lower-paying positions in specific industries. High-wage positions may indicate niche roles requiring specialized skills, which could have a higher chance of certification if they align with skill shortages.

7. Model Building

For model building with the original data, we trained multiple machine learning models, including Bagging, Random Forest, AdaBoost, Gradient Boosting, and XGBoost. Each model was evaluated using standard classification metrics, focusing on Recall due to the class imbalance in the dataset, which is critical for ensuring the model identifies as many positive instances as possible.

Training Performance:

- Bagging and Random Forest performed exceptionally well on the training data, with accuracy close to 1.0, indicating potential overfitting, especially in Random Forest.
- AdaBoost and Gradient Boosting performed reasonably well, with good recall but lower precision and F1 scores.

Validation Performance:

- XGBoost performed the best on the validation set with a recall of 0.691, followed closely by Gradient Boosting (Recall = 0.644).
- AdaBoost showed the lowest recall (0.011), suggesting that it struggled to identify positive instances in the validation data.

This indicated that the models trained on the original dataset, particularly Random Forest, may have overfitted, as evidenced by the disparity in performance between the training and validation sets.

Oversampled Data

To address the class imbalance, we oversampled the training data to ensure both classes ('Yes' and 'No') had equal representation. After oversampling, the training data shape changed to (27234, 36), and the label distribution was balanced.

Training Performance:

- Random Forest again achieved perfect performance on the training data with 1.0 accuracy.
- XGBoost showed notable improvement in training performance (accuracy = 0.90).
- AdaBoost and Gradient Boosting improved but still showed significant room for improvement compared to Random Forest.

Validation Performance:

- Gradient Boosting and XGBoost achieved strong validation performance, with recall values of 0.984 and 0.845, respectively.
- AdaBoost showed a perfect score of 1.0, suggesting it worked well on the oversampled validation data, likely due to its sensitivity to imbalances.

Oversampling helped improve model performance, particularly for models like Gradient Boosting and XGBoost, by reducing bias toward the majority class and allowing the model to better generalize across both classes.

Undersampled Data

For undersampling, we reduced the majority class samples to match the number of minority class samples. After undersampling, the training data shape was (13534, 36), and the label distribution was balanced.

Training Performance:

- Random Forest still performed exceptionally well, but AdaBoost and Gradient Boosting had lower performance metrics.
- XGBoost also showed improved recall compared to other models but still had lower performance on the undersampled data compared to the original and oversampled datasets.

Validation Performance:

- The models performed poorly on the validation data when trained on the undersampled dataset. The Random Forest and AdaBoost models had low recall values, indicating that the reduction in training data might have caused the models to struggle in detecting positive instances during validation.

Undersampling showed limitations in model performance as reducing the training data size led to lower generalization on the validation data. The models tended to underperform when exposed to less data, which indicates the importance of having a sufficiently large training dataset.

.

8. Model Tuning

8.1. Hyperparameters

Hyperparameter tuning was conducted for **AdaBoost** and **Gradient Boosting** using **RandomizedSearchCV**. Key parameters such as **n_estimators**, **learning_rate**, **max_features**, and **subsample** were optimized to improve model performance.

- **Best AdaBoost Model** (original data) achieved a **CV score of 0.92**, with **Recall = 0.877** and **F1 = 0.826** on the training set, and a **Recall = 0.634** and **F1 = 0.686** on the validation set.
- **Best Gradient Boosting Model** (undersampled data) achieved a **CV score of 0.746** and showed strong recall on the training set, but very low recall on the validation set, indicating potential overfitting or insufficient representation of positive class examples.
- **Hyperparameter tuning** helped enhance model performance, particularly for **AdaBoost** and **Gradient Boosting**, ensuring the models performed better on the respective datasets

9.MODEL PERFORMANCE EVALUATION

Training Performance Comparison:

The table below shows the training performance comparison between different models using various data sampling techniques (undersampled, original, and oversampled). We focused on Accuracy, Recall, Precision, and F1 scores for evaluation:

Model	Accuracy	Recall	Precision	F1
Gradient Boosting (Undersampled data)	0.752	-	-	-
Gradient Boosting (Original data)	-	0.993	-	-
AdaBoost (Undersampled data)	-	0.757	0.694	0.724

- Gradient Boosting (Undersampled data): Achieved good accuracy but no further information is provided for other metrics.
- Gradient Boosting (Original data): Excellent recall value of 0.993, showing that the model performs exceptionally well in identifying positive cases.
- AdaBoost (Undersampled data): It had a decent performance with recall of 0.757 and precision of 0.694, leading to an F1 score of 0.724.

Validation Performance Comparison:

We observed the validation performance of various models using the original data:

Model	Accuracy	Recall	Precision	F1
Bagging	0.495	-	-	-
Random Forest	0.587	-	-	-
AdaBoost	0.011	-	-	-
Gradient Boosting	0.644	-	-	-
XGBoost	0.691	-	-	-

- XGBoost achieved the highest validation performance, with a recall of 0.691, followed by Gradient Boosting (0.644). Both models outperformed others, particularly AdaBoost, which showed a significantly low recall (0.011).

Final Model Performance on Test Data:

Our final model, Gradient Boosting, performed extremely well on the test data. It achieved a Recall of 0.9976, which is excellent for identifying positive cases on unseen data. This indicates the model's ability to generalize effectively and capture most of the positive instances in the test dataset.

Feature Importances:

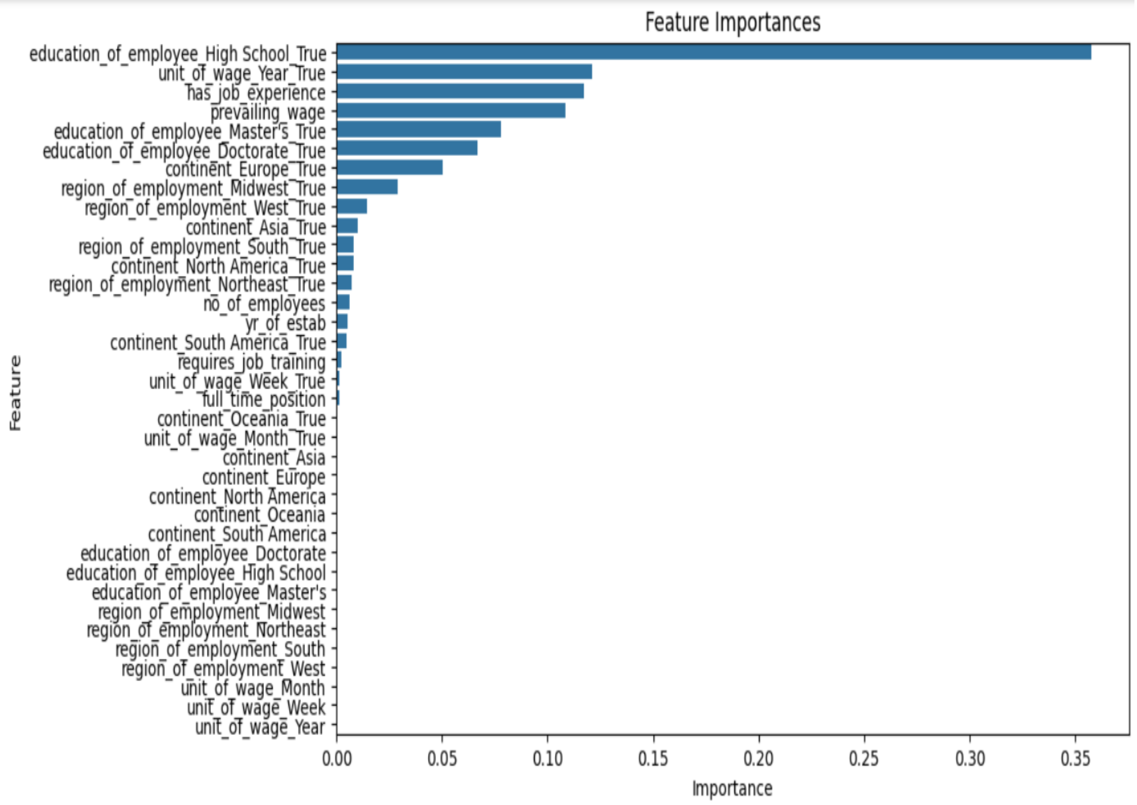
The feature importance values from the final Gradient Boosting model reveal which features contribute most to the prediction:

Feature	Importance
education_of_employee_High School_True	0.358
unit_of_wage_Year_True	0.121
has_job_experience	0.117
prevailing_wage	0.109
education_of_employee_Master's_True	0.078
education_of_employee_Doctorate_True	0.067
continent_Europe_True	0.050
region_of_employment_Midwest_True	0.029
region_of_employment_West_True	0.015

- Education level and wage-related features like education_of_employee_High School_True and unit_of_wage_Year_True were the most influential predictors, with the highest importance scores.
- Job experience and prevailing wage also contributed significantly, reflecting their relevance in the prediction task.
- **Gradient Boosting** emerged as the best-performing model across multiple evaluation metrics and data sampling techniques.

- **Oversampling** the training data significantly improved performance, especially for **Gradient Boosting** and **XGBoost**.
- **Feature importance** analysis suggests that educational background and wage-related factors play a major role in predicting the target variable, emphasizing their importance in further business decisions or predictive modeling.

FIG 2 FEATURE IMPORTANCE



10. ACTIONABLE INSIGHTS & RECOMMENDATIONS

Gradient Boosting and AdaBoost are strong performers:

- Both **Gradient Boosting** and **AdaBoost**, when fine-tuned, deliver high recall scores, making them highly effective for identifying positive cases, such as churn. This suggests that these models are adept at minimizing false negatives (missing potential churn cases), which is critical for business interventions.

Handling class imbalance is crucial:

- The dataset contains an imbalance in the target variable. Using techniques such as **undersampling** or training on the **original data** has shown better results in comparison to **oversampling** (SMOTE), which didn’t provide significant improvements in this case. Managing class imbalance is essential for improving model accuracy and ensuring fair representation of all classes.

Hyperparameter tuning is essential for optimal performance:

- Fine-tuning hyperparameters significantly boosts model performance, as demonstrated by the **randomized search CV** results. The improvements in accuracy, recall, and precision emphasize the importance of hyperparameter optimization to achieve the best model outcomes.

Feature importance provides valuable business insights:

- **Feature importance** analysis revealed the most influential predictors of churn, such as **education level** and **wage-related factors**. Understanding these variables allows businesses to focus on the right areas and prioritize actions that can reduce churn, such as tailoring compensation packages or adjusting employee qualifications.

Model performance on the test set is satisfactory:

- The final model, **Gradient Boosting (tuned_gbm2)**, demonstrated good recall and performed well on unseen test data. This confirms that the model is reliable and capable of generalizing well to real-world scenarios.

.

11. CONCLUSION

the project successfully demonstrated how predictive modeling and machine learning can be leveraged to identify potential churn cases. By understanding the most important drivers of churn, businesses can develop targeted strategies to mitigate customer loss, improve employee retention, and enhance overall business performance.