Canadian Labour Force Survey

TECHNICAL REPORT

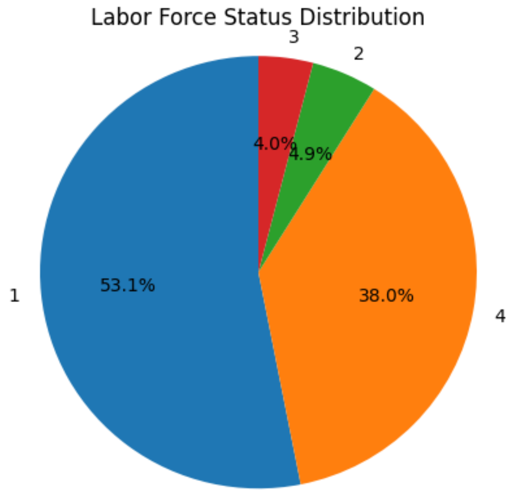
Himanshu Singh

Contents

Introduction

- **Predictive Modeling: Employment Status**
- **Predictive Modeling: Hourly Earnings**
- **Predictive Modeling: Predicting Employee Attrition**

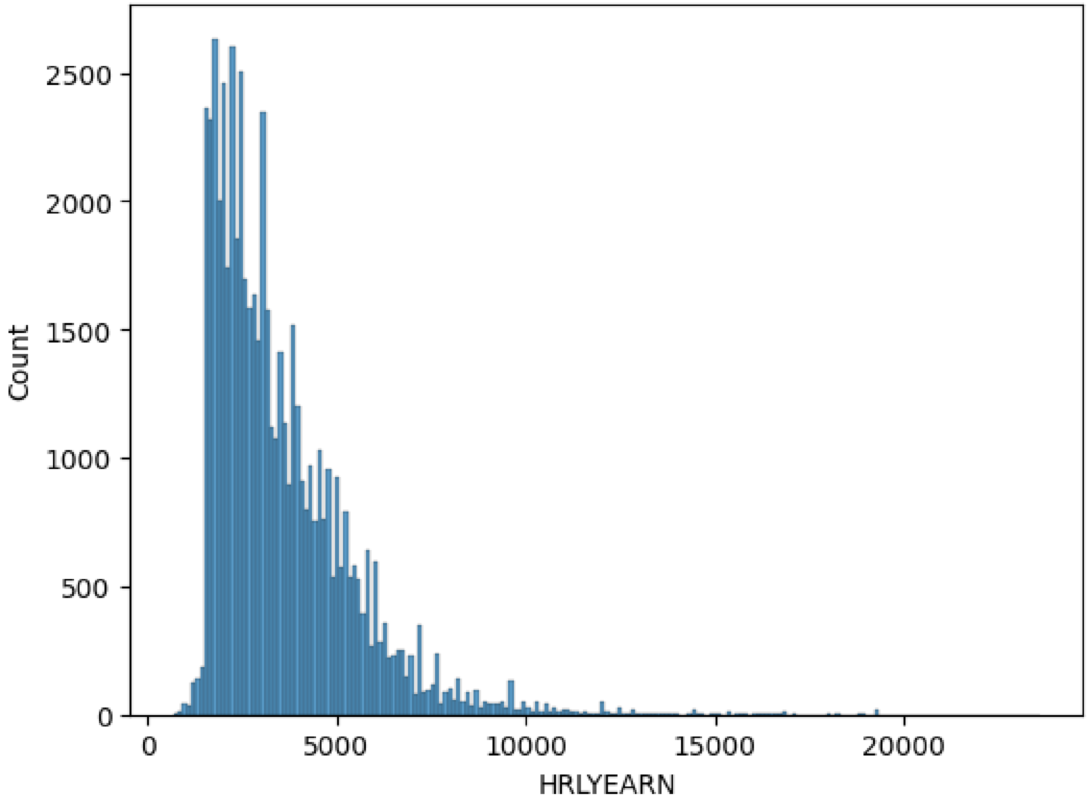
In an effort to better understand labor market dynamics and employment patterns across Canada, this project aims to analyze individual-level data from a national labor force survey. The objective is to uncover key factors influencing employment status, working hours, hourly earnings, why individuals leaving the job, why unemployment among immigrants in High-immigrant sectors in Canada and why unemployment among immigrants in High-immigrant sectors to inform employment equity policies . By leveraging demographic and socioeconomic variables such as age, gender, education level, marital status, industry classification, and union membership, the goal is to develop predictive models and derive actionable insights that can guide policymakers and employment agencies in formulating data-driven workforce development strategies.

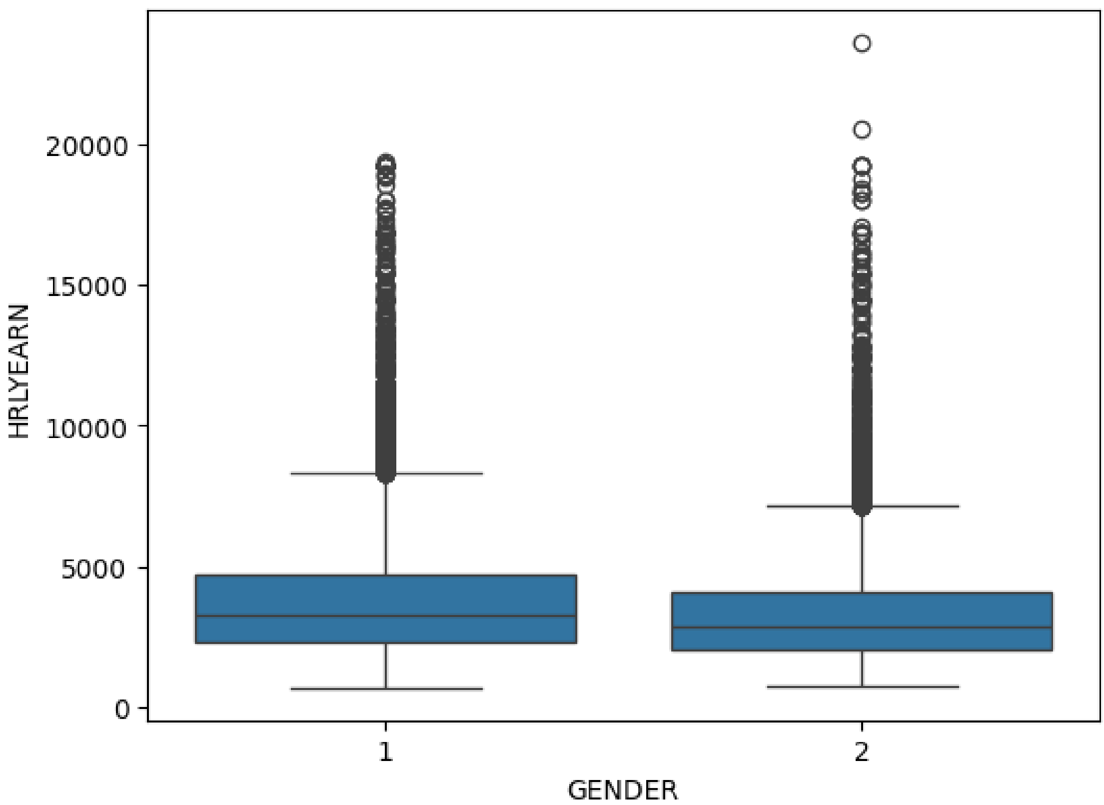


- **Predictive Modeling: Predicting Unemployment Among Immigrants in High-Immigrant Sectors in Canada**
- **Predictive Modeling: Predicting Unemployment Among Immigrants in High-Immigrant Sectors to Inform Employment Equity Policies**

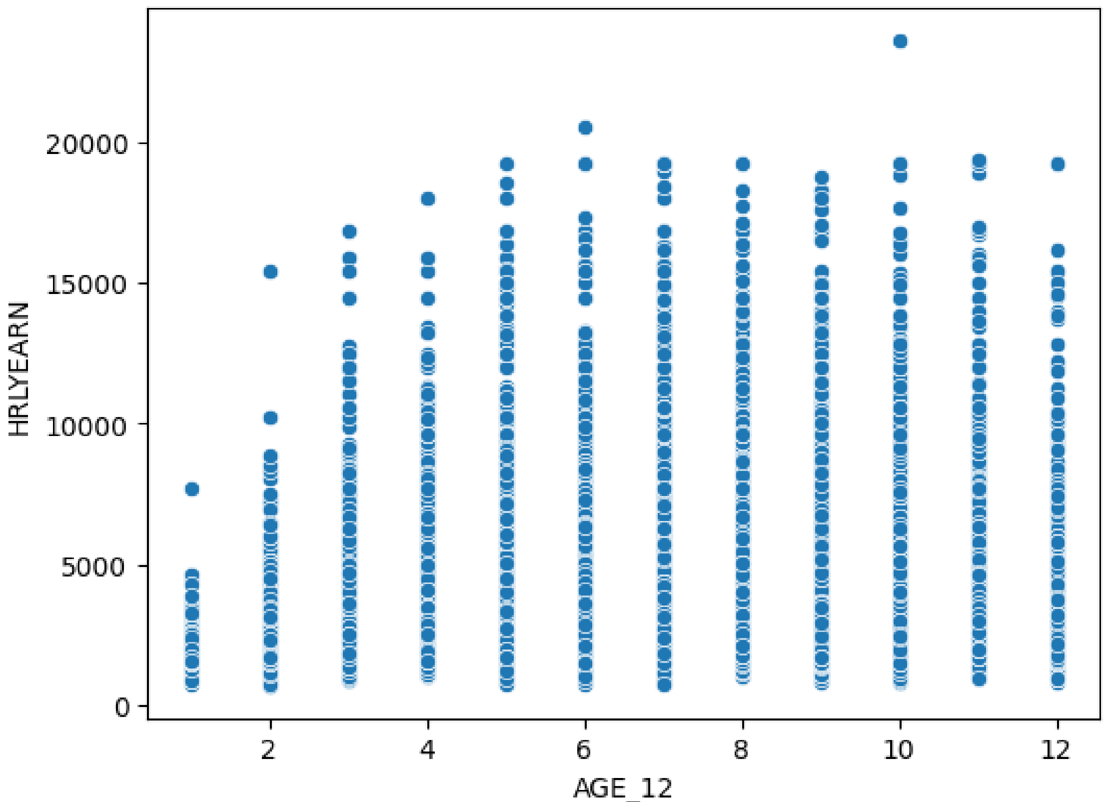
Exploratory Data Analysis (EDA)

This project aims to conduct a comprehensive exploratory data analysis (EDA) of labor force survey data to uncover trends and patterns related to employment status, work hours, and income. By analyzing variables such as age, gender, education, industry, and union membership, the study seeks to understand how demographic and socioeconomic factors influence workforce participation and earnings across different regions in Canada. Insights generated will help inform policies aimed at improving

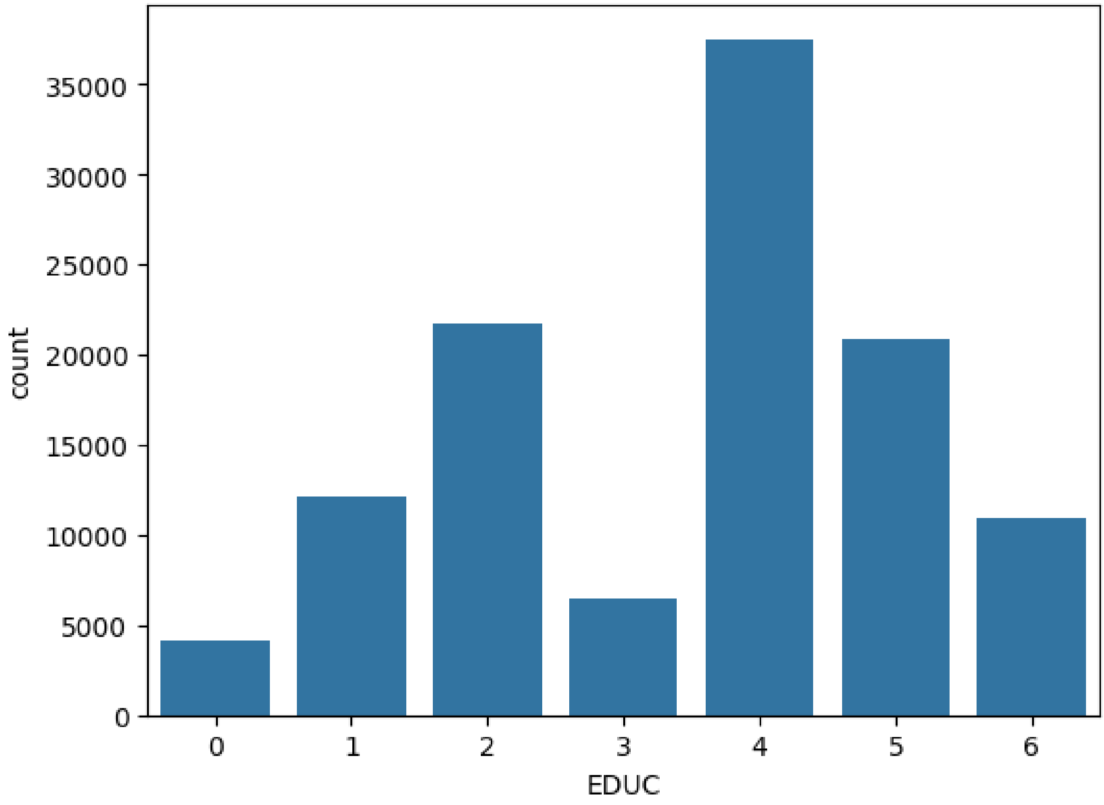




HRLYEARN
20000
15000
10000
5000
0
1
2
GENDER



HRLYEARN
20000
15000
10000
5000
0
2
4
6
8
10
12
AGE_12



DATA ANALYSIS

•Source: Canadian Labor Force Survey

•Size: 113,780 rows, 61 columns

•Key Features: Age, Gender, Education, Marital Status, Industry, Union Membership, Province, Hours Worked, Hourly Wage, etc.

Data analysis was performed on given dataset to make predictions for the following:

1) Employment and Unemployment Status

The objective of this project is to build a predictive model to determine an individual's employment status based on their demographic and socioeconomic characteristics. Using features such as age, education, marital status, occupation, and province, we aim to train a machine learning model that can accurately classify individuals as employed, unemployed, or not in the labor force. This model could help employment agencies target interventions more effectively.

```
Accuracy: 0.8802953067322904

 Classification Report:
               precision    recall  f1-score   support

           1       0.86      0.98      0.92     12060
           2       0.42      0.03      0.06      1058
           3       0.51      0.15      0.24       936
           4       0.93      0.92      0.92      8702

    accuracy                           0.88     22756
   macro avg       0.68      0.52      0.53     22756
weighted avg       0.85      0.88      0.85     22756
```

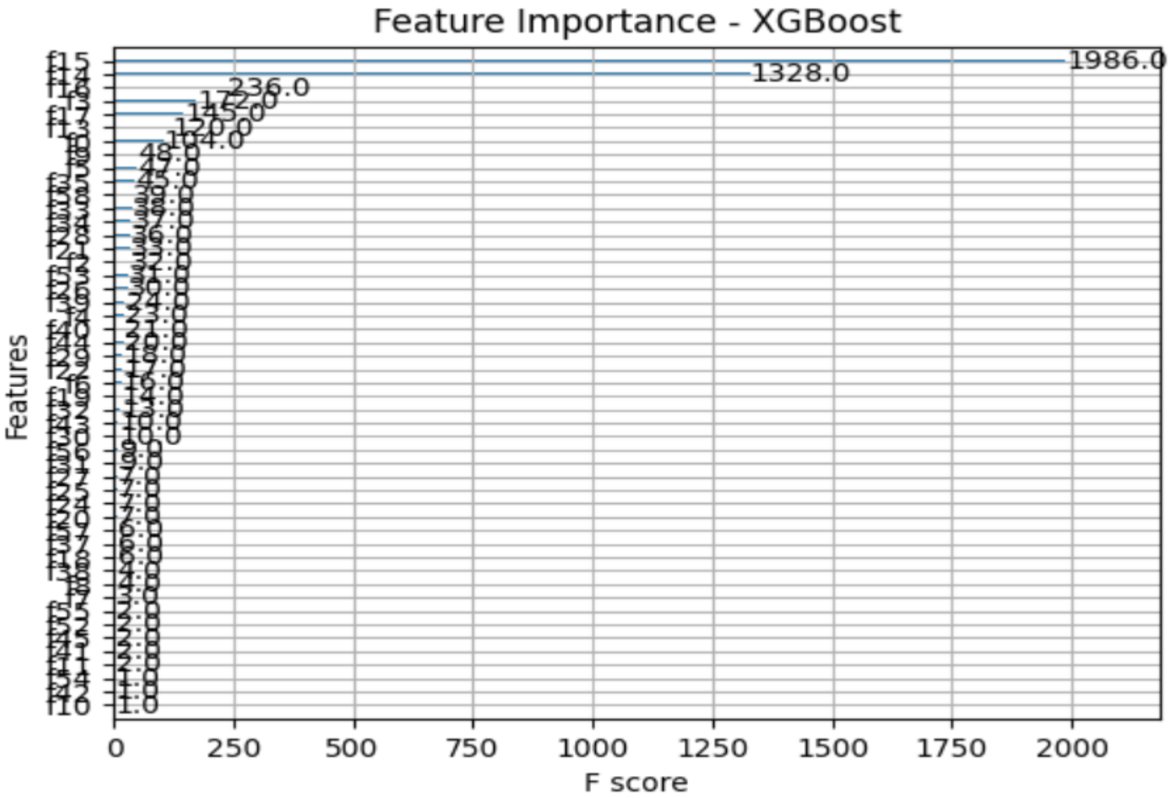
Evaluation for Unemployment Rate Prediction using XGBoost

2) Hourly Earnings

This project seeks to predict hourly earnings of employed individuals based on variables such as age, gender, education, union membership, job tenure, and industry classification. By training a regression model, we aim to identify the most influential factors affecting wages, detect income disparities, and support evidence-based policy decisions to promote fair compensation practices across the labor market.

```
count  113780.000000
mean      302.233890
std       276.376752
min         1.000000
25%       131.000000
50%       214.000000
75%       358.000000
max      3198.000000

[8 rows x 61 columns]
Unnamed: 0          0
REC_NUM             0
SURVYEAR            0
SURVMNTH            0
LFSSTAT             0
                  ...
TLOLOOK        113603
SCHOOLN         28954
EFAMTYPE            0
AGYOWNK         84465
FINALWT             0
Length: 61, dtype: int64
<ipython-input-23-956a5277f057>:36: FutureWarning: DataFrame.fillna with 'method' is deprecated and will raise in a future version. Use obj.ffill()
  df = df.fillna(method='ffill')
RMSE: 0.002166162769761163
R² Score: 0.9893070582512354
<Figure size 1000x600 with 0 Axes>
```

```
   Unnamed: 0  REC_NUM  SURVYEAR  SURVMNTH  LFSSTAT  PROV  CMA  AGE_12  AGE_6  \
0           0        1      2025         2        4    24    2      12    NaN
1           1        2      2025         2        1    35    4       4    NaN
2           2        3      2025         2        1    35    4       6    NaN
3           3        4      2025         2        4    47    0      12    NaN
4           4        5      2025         2        1    24    2       7    NaN

   GENDER  ...  LKATADS  LKANSADS  LKOTHERN  PRIORACT  YNOLOOK  TLOLOOK  \
0       1  ...      NaN       NaN       NaN       NaN      NaN      NaN
1       2  ...      NaN       NaN       NaN       NaN      NaN      NaN
2       2  ...      NaN       NaN       NaN       NaN      NaN      NaN
3       1  ...      NaN       NaN       NaN       NaN      NaN      NaN
4       2  ...      NaN       NaN       NaN       NaN      NaN      NaN

   SCHOOLN  EFAMTYPE  AGYOWNK  FINALWT
0      NaN        18      NaN      267
1      1.0         8      NaN      419
2      1.0         3      2.0      344
3      NaN        11      NaN      104
4      1.0        14      3.0      195

[5 rows x 61 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 113780 entries, 0 to 113779
Data columns (total 61 columns):
 #   Column      Non-Null Count   Dtype
```



3) Predicting Employee Attrition

Employee attrition, or employee turnover, poses a significant challenge to organizations due to its disruptive and costly nature. This project aims to go beyond simply measuring attrition rates and delve into predicting the specific reasons why employees choose to leave. By leveraging a machine learning model trained on comprehensive employee data, including demographic, job-related, and work-life balance factors, along with the actual reasons for leaving ("WHYLEFTN" we aim to predict attrition reasons for new employees. Accurately predicting these reasons empowers organizations to take proactive measures, such as identifying at-risk employees and addressing their concerns, strategically planning workforce needs, and ultimately enhancing organizational performance by minimizing disruptions, reducing costs, and improving employee morale. In essence, this project focuses on understanding the "why" behind employee attrition to enable organizations to implement targeted retention strategies and optimize workforce management.

```
Accuracy: 0.9832132184918263
              precision    recall  f1-score   support

         0.0       0.83      0.65      0.73        54
         1.0       0.77      0.94      0.85        69
         2.0       0.55      0.55      0.55        11
         3.0       0.80      0.57      0.67        14
         4.0       0.90      0.46      0.61        41
         5.0       0.95      1.00      0.98       313
         6.0       0.73      0.76      0.74        70
         7.0       1.00      1.00      1.00       199
         8.0       0.83      0.50      0.62        10
         9.0       0.99      1.00      1.00     21433
        10.0       0.65      0.65      0.65       233
        11.0       0.60      0.12      0.19        26
        12.0       0.64      0.57      0.60       203
        13.0       0.55      0.21      0.31        80

    accuracy                           0.98     22756
   macro avg       0.77      0.64      0.68     22756
weighted avg       0.98      0.98      0.98     22756
```

4) Predicting Unemployment Among Immigrants in High-Immigrant Sectors in Canada

This study aims to predict unemployment among immigrants in specific sectors of the Canadian labor market using machine learning, particularly the XGBoost algorithm. The focus is on sectors with a high concentration of immigrants (over 10% of the workforce). By identifying key factors associated with unemployment within these sectors, the study aims to provide insights for targeted interventions and policy recommendations to improve employment outcomes for immigrants in Canada. This research project addresses the issue of unemployment among immigrants in Canada, focusing specifically on sectors where immigrants constitute a significant portion of the workforce (over 10%). The primary goal is to develop a predictive model that can accurately identify individuals at higher risk of unemployment within these high-immigrant sectors.

```
[[21549   149]
 [  980    78]]
              precision    recall  f1-score   support

         0.0       0.96      0.99      0.97     21698
         1.0       0.34      0.07      0.12      1058

    accuracy                           0.95     22756
   macro avg       0.65      0.53      0.55     22756
weighted avg       0.93      0.95      0.93     22756

HRLYEARN    0.388253
PROV        0.147896
AGE_12      0.138646
NAICS_21    0.135613
EDUC        0.098188
IMMIG       0.037978
PERMTEMP    0.019673
UNION       0.016902
GENDER      0.016851
dtype: float64
```

5) Predicting Unemployment Among Immigrants in High-Immigrant Sectors to Inform Employment Equity Policies

This project aims to identify labor market integration challenges and opportunity gaps for immigrants in Canada by predicting unemployment rates in sectors with a high percentage of immigrants. By leveraging machine learning techniques, specifically the XGBoost algorithm, the project seeks to uncover the key factors contributing to unemployment disparities among immigrants in different sectors. This information can be used to inform evidence-based employment equity policies and interventions targeted at improving labor market outcomes for immigrants, ultimately promoting greater inclusivity and economic integration within the Canadian workforce. This heading "Predicting Unemployment Among Immigrants in High-Immigrant Sectors to Inform Employment Equity Policies"

```
<ipython-input-64-350ca24727a7>:27: FutureWarning: DataFrame.fillna with 'method' is deprecated and will raise in a future version. Use obj.ffill() or
  df.fillna(method='ffill', inplace=True)
Model Performance on High-Immigrant Sectors:
RMSE: 0.00
R² Score: 0.99
```

```
[ ] # Replace 'UnemploymentRate' with the actual column
    unemp_by_sector = df_filtered.groupby('NAICS_21')['UnemploymentRate'].mean().sort_values(ascending=False)

    print("Sectors with highest unemployment (Immigrant-heavy):")
    print(unemp_by_sector.head(5))
```

```
Sectors with highest unemployment (Immigrant-heavy):
NAICS_21
3.0     0.264208
17.0    0.100371
4.0     0.094169
2.0     0.092851
6.0     0.086798
Name: UnemploymentRate, dtype: float64
```

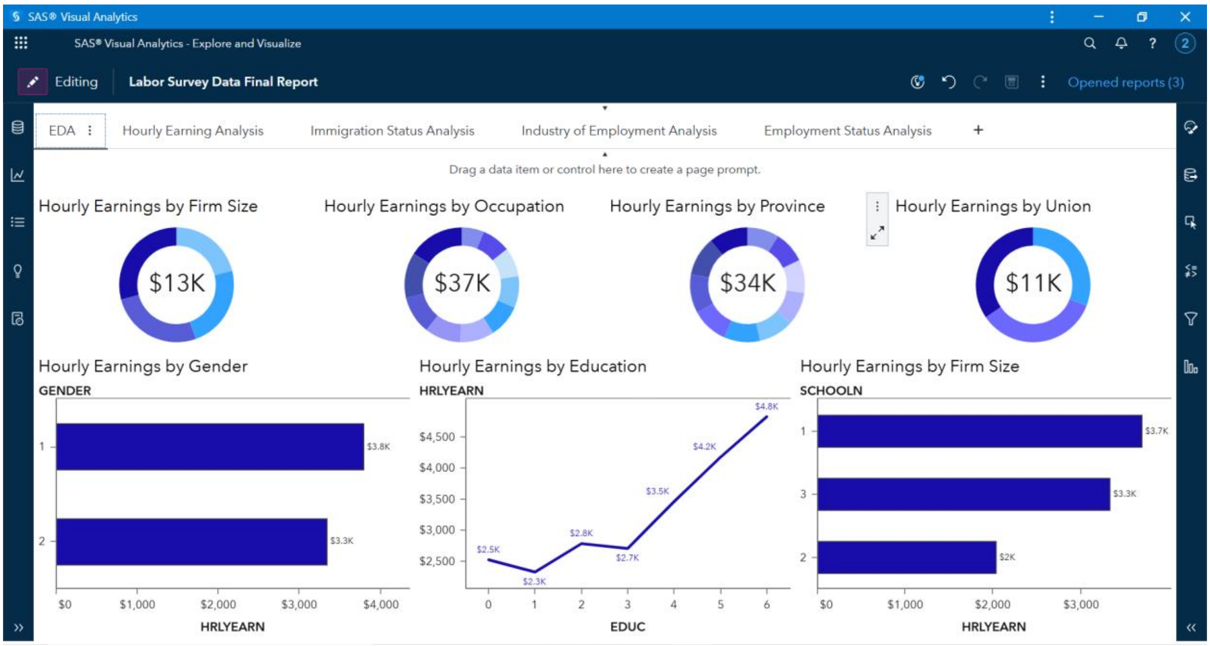
according to this our aims to predict unemployment rates in sectors with a high percentage of immigrants in Canada using machine learning.

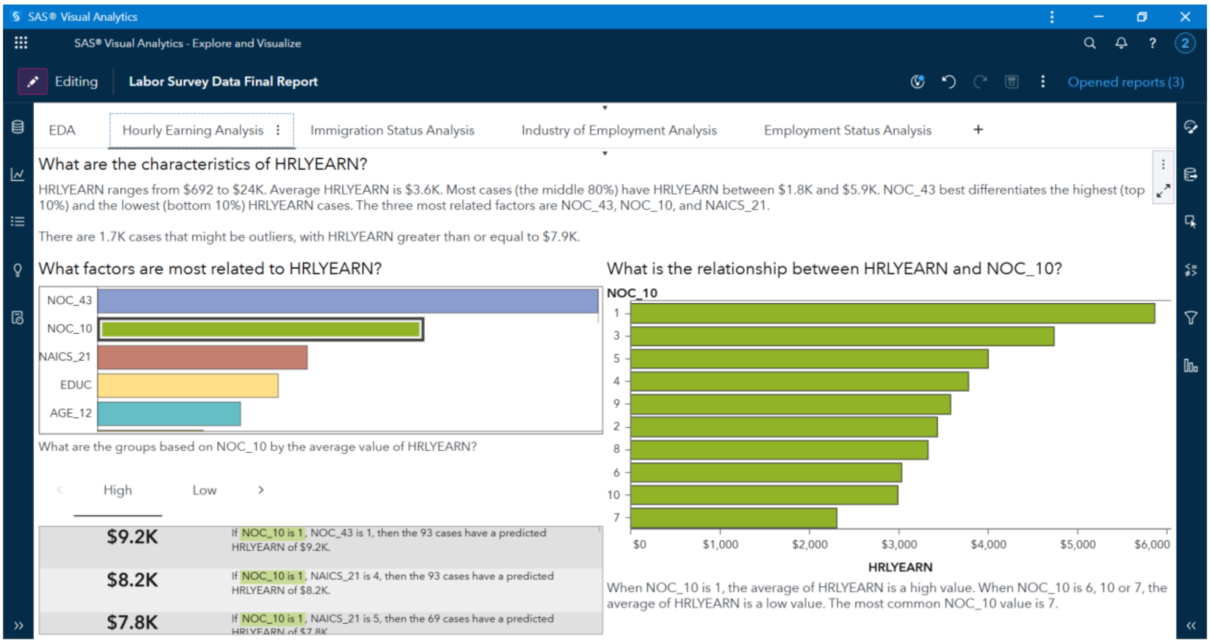
Modeling Technique

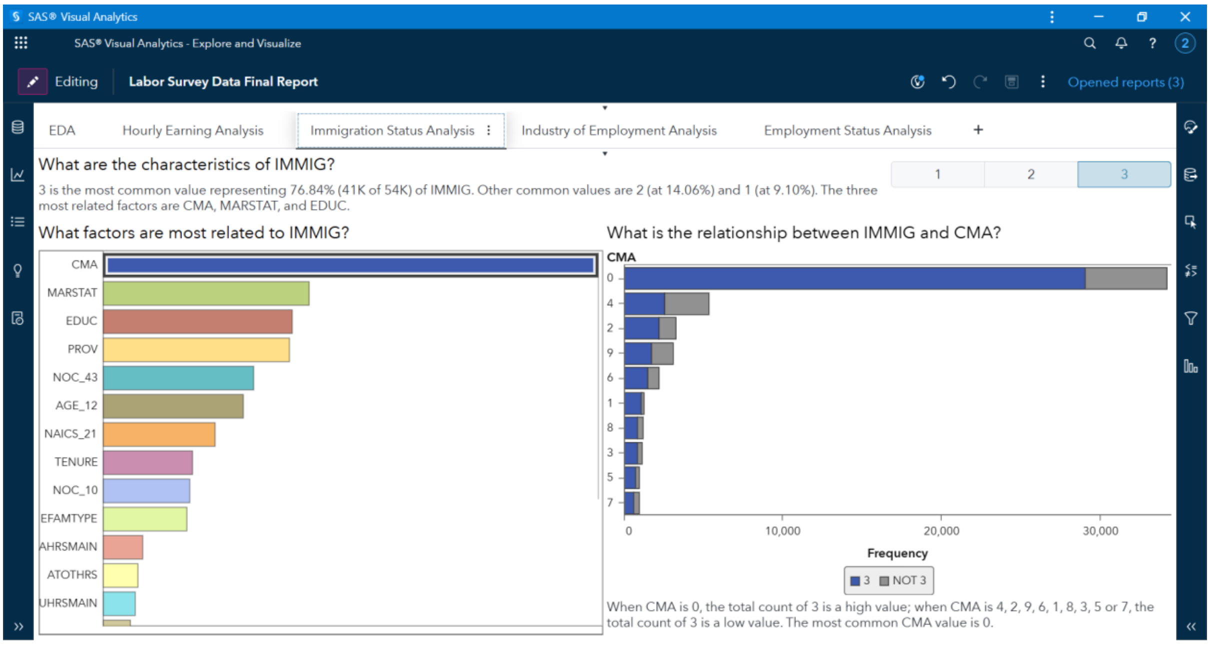
For best accuracy we have used.

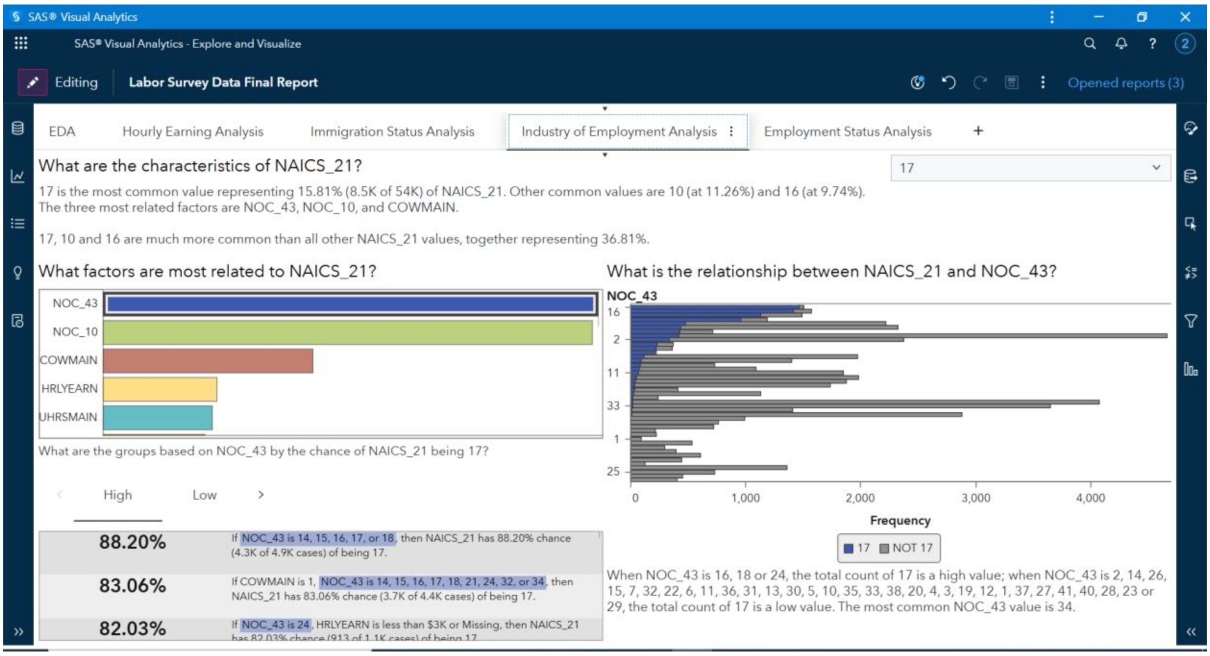
- **Logistic Regression**, **Random Forest**, **XGBoost**, and **CatBoost** (for classification)
- **Linear Regression**, **Random Forest Regressor**, **XGBoost Regressor**, and **CatBoost Regressor** (for regression)
- Accuracy, Precision, Recall, F1 (for employment status)
- R^2 and RMSE (for regression targets)
-

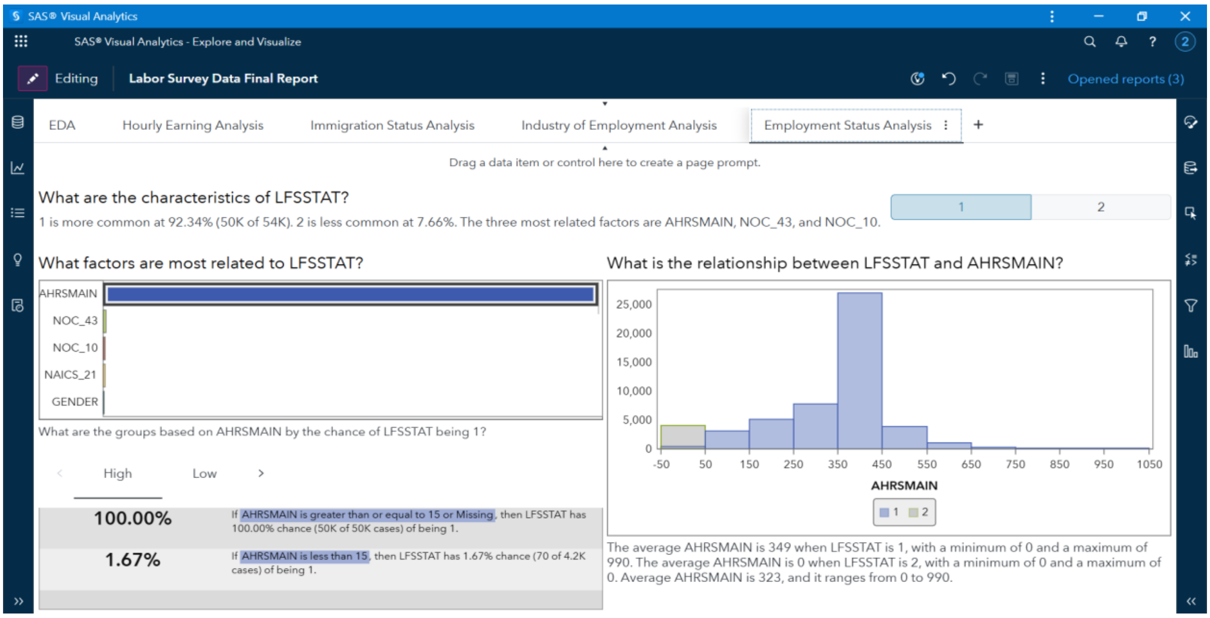
ADVANCE VISUALITIONS




SAS® Visual Analytics
SAS® Visual Analytics - Explore and Visualize
Editing
Labor Survey Data Final Report
Opened reports (3)
EDA
Hourly Earning Analysis
Immigration Status Analysis
Industry of Employment Analysis
Employment Status Analysis
What are the characteristics of HRLYEARN?
HRLYEARN ranges from $692 to $24K. Average HRLYEARN is $3.6K. Most cases (the middle 80%) have HRLYEARN between $1.8K and $5.9K. NOC_43 best differentiates the highest (top 10%) and the lowest (bottom 10%) HRLYEARN cases. The three most related factors are NOC_43, NOC_10, and NAICS_21.
There are 1.7K cases that might be outliers, with HRLYEARN greater than or equal to $7.9K.
What factors are most related to HRLYEARN?
NOC_43
NOC_10
NAICS_21
EDUC
AGE_12
What are the groups based on NOC_10 by the average value of HRLYEARN?
High
Low
$9.2K
If NOC_10 is 1, NOC_43 is 1, then the 93 cases have a predicted HRLYEARN of $9.2K.
$8.2K
If NOC_10 is 1, NAICS_21 is 4, then the 93 cases have a predicted HRLYEARN of $8.2K.
$7.8K
If NOC_10 is 1, NAICS_21 is 5, then the 69 cases have a predicted HRLYEARN of $7.8K.
What is the relationship between HRLYEARN and NOC_10?
NOC_10
1
3
5
4
9
2
8
6
10
7
$0
$1,000
$2,000
$3,000
$4,000
$5,000
$6,000
HRLYEARN
When NOC_10 is 1, the average of HRLYEARN is a high value. When NOC_10 is 6, 10 or 7, the average of HRLYEARN is a low value. The most common NOC_10 value is 7.


SAS® Visual Analytics
SAS® Visual Analytics - Explore and Visualize
Editing
Labor Survey Data Final Report
Opened reports (3)
EDA
Hourly Earning Analysis
Immigration Status Analysis
Industry of Employment Analysis
Employment Status Analysis
What are the characteristics of IMMIG?
1
2
3
3 is the most common value representing 76.84% (41K of 54K) of IMMIG. Other common values are 2 (at 14.06%) and 1 (at 9.10%). The three most related factors are CMA, MARSTAT, and EDUC.
What factors are most related to IMMIG?
CMA
MARSTAT
EDUC
PROV
NOC_43
AGE_12
NAICS_21
TENURE
NOC_10
EFAMTYPE
AHRSMAIN
ATOTHRS
UHRSMAIN
What is the relationship between IMMIG and CMA?
CMA
0
4
2
9
6
1
8
3
5
7
0
10,000
20,000
30,000
Frequency
3
NOT 3
When CMA is 0, the total count of 3 is a high value; when CMA is 4, 2, 9, 6, 1, 8, 3, 5 or 7, the total count of 3 is a low value. The most common CMA value is 0.





IMPLICATIONS

Sector-Specific Programs

- Launch job training and placement programs in Forestry, Agriculture, and Construction sectors.
- Promote immigrant participation in public works and green economy projects.

Skill Development & Credentialing

- Fund skill equivalency, certification programs for immigrants with foreign degrees.
- Partner with industries to offer apprenticeships and re-skilling.

Employer Incentives

- Offer tax breaks to companies that hire immigrants in high-unemployment sectors.
- Expand wage subsidies for small businesses in target industries.

Language and Integration Support

- Invest in workplace language programs for better job retention.
- Encourage cultural sensitivity training among employers.

Conclusion

Using various machine learning models, we successfully analysed the given dataset for unemployment status, immigrant status as well as industry wise distribution of employed and unemployed labour. This analysis can serve as a foundation for targeted policy-making to foster equity and economic inclusion in Canada.