April 14, 2023

TERM PROJECT REPORT

Predicting Income Levels Using Machine Learning Techniques

Presented to

Ivan Wong

Presented by

Nhi Cao - 300 367 933

Amrit Sian - 300 340 252

Huy Thuy Dung Nguyen - 300 363 745

TABLE OF CONTENTS

INTRODUCTION AND DISCOVERY

Business Domain

The goal of this project is to determine whether a person makes over 50K a year using the Adult dataset available from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Adult). The dataset contains various demographic, work, and financial attributes, which can be used to develop a predictive model. Identifying individuals who make more than 50K a year can help organizations in targeted marketing, policy planning, and social welfare program allocation.

Framing the Problem

This project is essential for several reasons:

- Identifying factors that contribute to higher income can help individuals make better career and educational choices.
- Governments and policymakers can use the insights to develop more effective social welfare programs and policies that target specific population segments.
- Companies and organizations can benefit from understanding the factors that drive income levels when creating their compensation policies and recruitment strategies.

INTRODUCTION AND DISCOVERY

Initial Hypotheses

- Age is positively correlated with income. Older individuals might have more work experience and higher positions, leading to higher salaries.
- Higher education levels will be associated with higher income levels. People with advanced degrees might be more likely to have high-paying jobs.
- The number of hours worked per week will be positively correlated with income. Full-time workers might earn more than part-time workers.
- Occupation type will have a significant impact on income levels. Certain industries and roles might offer higher salaries than others.
- Marital status might influence income levels. Married individuals might have higher combined household incomes, leading to higher reported incomes per person.
- Gender might be a significant factor in determining income, given the known gender pay gap that exists in various industries.

As we analyze the dataset, we will test these hypotheses and possibly identify additional factors influencing individuals' income levels.

DATA PREPARATION

Data Inventory

The Adult dataset was obtained from the UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/Adult. This data was extracted from the census bureau database found at http://www.census.gov/ftp/pub/DES/www/welcome.html

Donor: Ronny Kohavi and Barry Becker, Data Mining and Visualization, Silicon Graphics.

The dataset contains 32,561 instances and 15 attributes. For this project, 20% of the data (6,512 instances) will be randomly sampled and set aside to make out-of-sample predictions. The target variable is 'income', which is a binary classification task.

1	age	workclass	fnlwgt	education	educati	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-	native-country	income
2	39	State-gov	77516	Bachelors	13	Never-marrie	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
3	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-s	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
4	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaner	Not-in-family	White	Male	0	0	40	United-States	<=50K
5	53	Private	234721	11th	7	Married-civ-s	Handlers-cleaner	Husband	Black	Male	0	0	40	United-States	<=50K
6	28	Private	338409	Bachelors	13	Married-civ-s	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
7	37	Private	284582	Masters	14	Married-civ-s	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
8	49	Private	160187	9th	5	Married-spou	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
9	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-s	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
10	31	Private	45781	Masters	14	Never-marrie	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
11	42	Private	159449	Bachelors	13	Married-civ-s	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
12	37	Private	280464	Some-college	10	Married-civ-s	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
13	30	State-gov	141297	Bachelors	13	Married-civ-s	Prof-specialty	Husband	Asian-Pac	Male	0	0	40	India	>50K
14	23	Private	122272	Bachelors	13	Never-marrie	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
15	32	Private	205019	Assoc-acdm	12	Never-marrie	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
16	40	Private	121772	Assoc-voc	11	Married-civ-s	Craft-repair	Husband	Asian-Pac	Male	0	0	40	?	>50K
17	34	Private	245487	7th-8th	4	Married-civ-s	Transport-moving	Husband	Amer-Ind	Male	0	0	45	Mexico	<=50K
18	25	Self-emp-not-inc	176756	HS-grad	9	Never-marrie	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
19	32	Private	186824	HS-grad	9	Never-marrie	Machine-op-inspc	Unmarried	White	Male	0	0	40	United-States	<=50K
20	38	Private	28887	11th	7	Married-civ-s	Sales	Husband	White	Male	0	0	50	United-States	<=50K
21	43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
22	40	Private	193524	Doctorate	16	Married-civ-s	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
23	54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
24	35	Federal-gov	76845	9th	5	Married-civ-s	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
25	43	Private	117037	11th	7	Married-civ-s	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
26	59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K
27	56	Local-gov	216851	Bachelors	13	Married-civ-s	Tech-support	Husband	White	Male	0	0	40	United-States	>50K
28	19	Private	168294	HS-grad	9	Never-marrie	Craft-repair	Own-child	White	Male	0	0	40	United-States	<=50K
29	54	?	180211	Some-college	10	Married-civ-s	?	Husband	Asian-Pac	Male	0	0	60	South	>50K
30	39	Private	367260	HS-grad	9	Divorced	Exec-managerial	Not-in-family	White	Male	0	0	80	United-States	<=50K

DATA PREPARATION

Data Inventory - Attribute Information

1. income: >50K, <=50K.
2. age: continuous.
3. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
4. fnlwgt: continuous.
5. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
6. education-num: continuous.
7. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
8. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
9. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
10. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
11. sex: Female, Male.
12. capital-gain: continuous.
13. capital-loss: continuous.
14. hours-per-week: continuous.
15. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

DATA PREPARATION

Data Processing

- Dropped 'education' (redundant) and 'fnlwgt' (irrelevant) columns
- Dropped 'native_country' column to reduce dimensionality
- Replaced '?' with NaN and dropped rows with missing values
- Dataset size after cleaning: 24,567 entries out of 26,049
- Created dummy variables for categorical features
- Final dataframe has 41 columns
- Generated correlation heatmap to identify strong correlations between features

```
df1.describe()
```

	age	education_num	capital_gain	capital_loss	hours_per_week	workclass_Local_gov	workclass_Private	workclass_Self_emp_inc	workclass_Self_emp_not_inc	workclass_State_gov	...
count	24567.000000	24567.000000	24567.000000	24567.000000	24567.000000	24567.000000	24567.000000	24567.000000	24567.000000	24567.000000	...
mean	38.454105	10.117963	1094.647902	88.626287	40.912240	0.067407	0.740017	0.036879	0.083079	0.041478	...
std	13.139511	2.559256	7394.750362	405.921531	11.961921	0.250732	0.438634	0.188468	0.276007	0.199398	...
min	17.000000	1.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
25%	28.000000	9.000000	0.000000	0.000000	40.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
50%	37.000000	10.000000	0.000000	0.000000	40.000000	0.000000	1.000000	0.000000	0.000000	0.000000	...
75%	47.000000	13.000000	0.000000	0.000000	45.000000	0.000000	1.000000	0.000000	0.000000	0.000000	...
max	90.000000	16.000000	99999.000000	4356.000000	99.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...

8 rows × 41 columns

```
df1.head(10)
```

	age	education_num	capital_gain	capital_loss	hours_per_week	workclass_Local_gov	workclass_Private	workclass_Self_emp_inc	workclass_Self_emp_not_inc	workclass_State_gov	...	relationship_Other_relative	relationship_Own_child	relationship_Unmarried
0	49	9	0	0	40	0	0	0	1	0	...	0	0	0
1	17	7	0	0	22	0	1	0	0	0	...	1	0	0
2	22	9	0	0	40	0	1	0	0	0	...	0	0	0
3	23	9	0	0	53	0	1	0	0	0	...	0	1	0
4	38	13	7688	0	55	0	1	0	0	0	...	0	0	0
5	32	14	0	0	38	0	0	0	0	1	...	0	0	0
6	48	9	7298	0	50	0	0	1	0	0	...	0	0	0
7	32	9	0	0	40	0	1	0	0	0	...	0	0	0
8	54	9	0	0	40	0	1	0	0	0	...	0	0	0
9	41	9	0	0	40	0	1	0	0	0	...	0	0	0

10 rows × 41 columns

MODEL PLANNING AND IMPLEMENTATION

Proposed models and Techniques

- Employed multiple classifiers for comparison:
  - Logistic Regression: simple, interpretable, and widely used for binary classification tasks
  - Random Forest: robust, handles high dimensional data, reduces overfitting
  - AdaBoost: adaptive boosting technique, improves base classifier performance
  - XGBoost: powerful gradient boosting algorithm, known for handling imbalanced data and delivering high accuracy
  - MLP (Multi-Layer Perceptron): flexible neural network, can capture complex patterns and nonlinear relationships
- Utilized ensemble techniques:
  - Voting Classifiers (hard and soft voting)
  - Bagging Classifier
- Incorporated feature selection and scaling methods:
  - SelectFromModel (using Random Forest as the estimator)
  - Recursive Feature Elimination (using Logistic Regression as the estimator)
  - MinMaxScaler
  - StandardScaler
- Integrated GridSearchCV for hyperparameter tuning:
  - Performed an initial search with a variety of classifier options
  - Identified the best combination (XGBClassifier with SelectFromModel and MinMaxScaler)
  - Performed a more focused search for optimal hyperparameters for the XGBClassifier

This approach allows for an efficient exploration of various models and techniques, ultimately leading to the selection of the most effective model for the given dataset.

MODEL PLANNING AND IMPLEMENTATION

Efficient Project Workflow

- Pipelines:
  - Streamline preprocessing, feature selection, and model fitting steps
  - Simplify code and reduce errors by applying a sequence of operations in a single object
- Feature selection methods:
  - SelectFromModel and Recursive Feature Elimination (RFE) help identify the most relevant features
  - Reduces model complexity and improves training efficiency
- Scaling techniques:
  - MinMaxScaler and StandardScaler preprocess the data to ensure all features have the same scale
  - Enhances model performance and speeds up training
- Cross-validation:
  - Enables assessment of model performance using different training and validation sets
  - Reduces the risk of overfitting and provides a better estimate of the model's generalization ability
- GridSearchCV:
  - Performs an exhaustive search for optimal hyperparameters
  - Automates the process of selecting the best combination of parameters to improve model performance
  - Saves time and resources by evaluating multiple models in parallel (using the n_jobs parameter)

MODEL PLANNING AND IMPLEMENTATION

Optimize Hypothesis Testing and Modeling Insights

- Model performance metrics (accuracy_score, classification_report, confusion_matrix, ROC curve):
  - Provide a comprehensive assessment of the model's performance, allowing for the evaluation of the hypotheses and modeling objectives.
  - Offer insights into the model's ability to predict income levels accurately and how well it can generalize to new data.

- Feature importances from XGBClassifier:
  - Quantify the relative importance of each feature in the model, allowing for assessment of the hypotheses and identification of the most significant factors influencing income.
  - Offer insights into which features have the strongest impact on the model's predictions and help validate or refute the initial hypotheses.

MODEL PLANNING AND IMPLEMENTATION

Optimize Hypothesis Testing and Modeling Insights

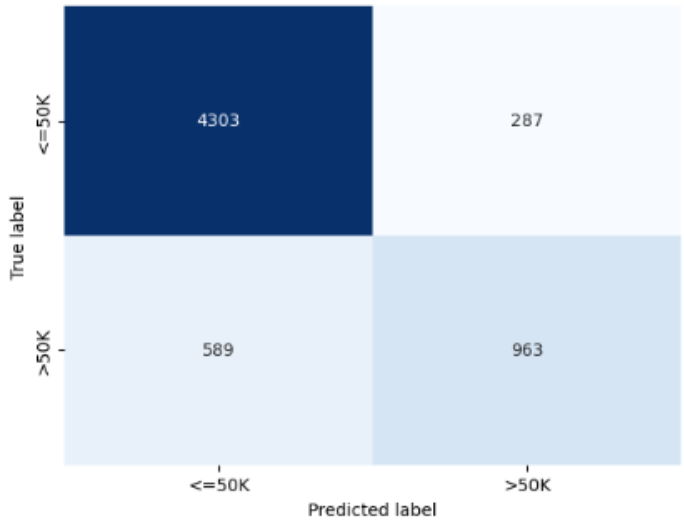
Feature	Importance
marital_status_Married_civ_spouse	0.642812
education_num	0.096745
capital_gain	0.087856
occupation_Exec_managerial	0.053241
capital_loss	0.047009
age	0.027213
hours_per_week	0.022747
sex_Male	0.022377

- Marital status (Married_civ_spouse) has the highest impact on income levels, supporting the hypothesis that marital status influences income.
- Education_num confirms that higher education levels are associated with higher income levels, with relatively high importance.
- Occupation (Exec_managerial) supports the hypothesis that occupation type impacts income levels, particularly for higher-paying roles.
- Age has a positive but weaker correlation with income levels compared to other factors, partially confirming the hypothesis that older individuals have higher salaries.
- Hours_per_week confirms the hypothesis that there is a correlation between hours worked and income levels, but with weaker importance than other factors.
- Sex (Male) indicates that gender is a factor in determining income levels, supporting the hypothesis that the gender pay gap affects income.
- Capital_gain and capital_loss, not part of the initial hypotheses, provide additional insight into factors affecting income levels.

RESULTS INTERPRETATION AND IMPLICATIONS

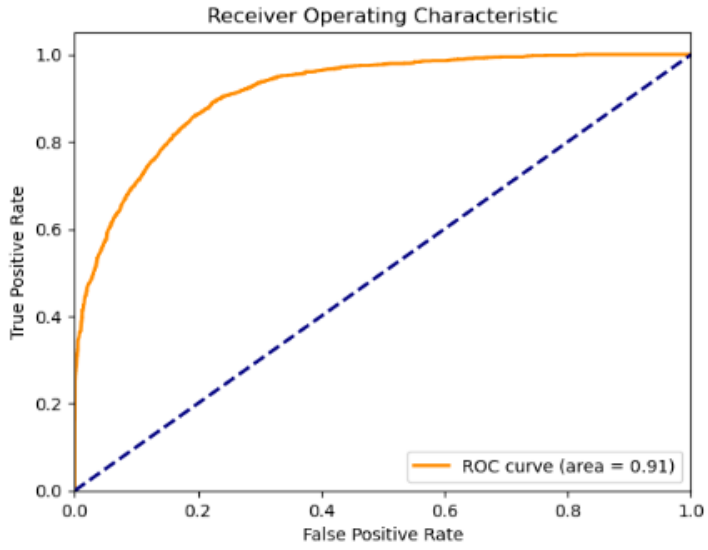
Summary of Results

Model accuracy: 0.857



	precision	recall	f1-score	support
0	0.88	0.94	0.91	4590
1	0.77	0.62	0.69	1552
accuracy			0.86	6142
macro avg	0.82	0.78	0.80	6142
weighted avg	0.85	0.86	0.85	6142

- Model accuracy: 0.857
- Confusion Matrix:
  - True Negative (<=50K): 4303
  - False Negative: 589
  - True Positive (>50K): 963
  - False Positive: 287



- ROC curve (area = 0.91)

RESULTS INTERPRETATION AND IMPLICATIONS

Assessment of Results

- Model Validity and Accuracy:
  - Performance metrics: Accuracy of 0.857 and weighted average F1-score of 0.85, indicating a reasonably good performance.
  - Confusion matrix: 4303 TN, 963 TP, 287 FP, and 589 FN, suggesting a conservative approach in predicting '>50K' class, with some misclassifications.
  - ROC-AUC curve: AUC of 0.91, suggesting good discriminative power of the model.
- Model Output/Behavior: The model's output and behavior seem reasonable and consistent with domain knowledge. The most important feature is marital status (specifically being married to a civilian spouse), followed by education, capital gain, and occupation.
- Parameter Values: The parameter values used in the model make sense in the context of the domain.
- Model Sufficiency: The model's accuracy of 85.7% suggests that it may be sufficiently accurate to meet the goal, depending on the specific requirements of the application.
- Model Limitation: The model could be improved in predicting the '>50K' class, as indicated by the precision and recall values for this class.
- More Data or Inputs: The current features and data seem to have resulted in a reasonably accurate model. However, gathering more data or exploring additional features could potentially further improve the model's performance.
- Model Form: The current form of the model, using an XGBClassifier and a pipeline with feature selection and scaling, appears to address the problem effectively. Alternative models could be explored.

RESULTS INTERPRETATION AND IMPLICATIONS

Assessment of Results

Does the model avoid intolerable mistakes?

In the confusion matrix, we have:

- False Positives: 287 - The model incorrectly predicts that an individual earns >50K when they actually earn <=50K.
- False Negatives: 589 - The model incorrectly predicts that an individual earns <=50K when they actually earn >50K.

The impact of these mistakes depends on the specific application of the model. For instance, if the model is used to **determine eligibility for financial aid or social benefits**, false positives might lead to individuals receiving aid they do not actually qualify for, while false negatives could prevent deserving individuals from receiving the assistance they need. In another example, if the model is used by a **marketing** team to target high-income customers, false positives could result in wasted marketing resources on lower-income individuals, while false negatives could lead to missed opportunities with potential high-income customers. Take credit scoring as a third example. If the model is used as a component in **credit scoring systems**, false positives might cause lenders to overestimate the creditworthiness of individuals with lower incomes, potentially increasing the risk of default, while false negatives could prevent higher-income individuals from accessing credit at favorable terms.

To determine if these mistakes are tolerable, the stakeholders and domain experts need to assess the cost and consequences of these errors in the context of their specific application. If the costs and consequences are deemed too high, further model refinement or alternative approaches may be necessary to minimize these errors and create a more accurate and reliable model for the intended use case.

RESULTS INTERPRETATION AND IMPLICATIONS

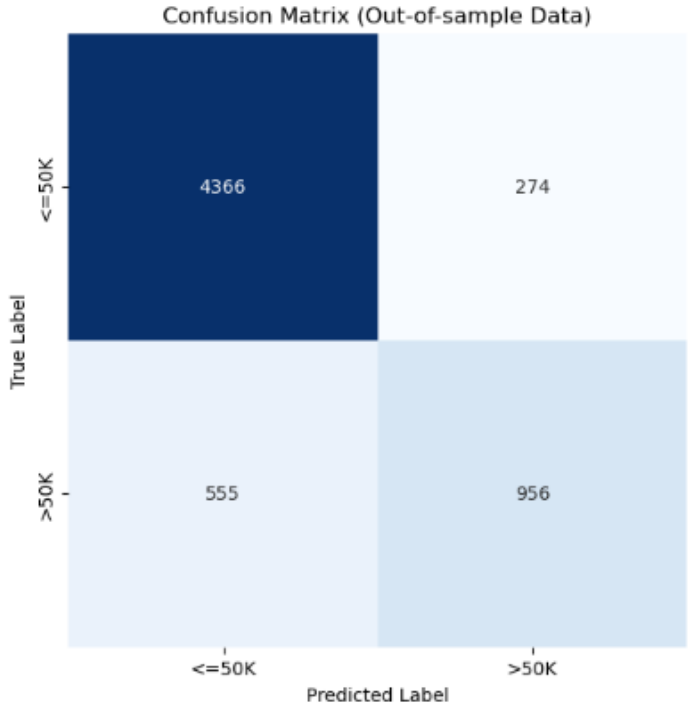
Key Findings and Major Insights

- Model performance: The XGBClassifier model achieved an accuracy of 0.857 on the test data, demonstrating a relatively good performance in predicting income levels.
- Precision and recall: The model has higher precision and recall for the '<=50K' class, while the '>50K' class could benefit from further improvements in the model's predictive ability.
- ROC curve: The ROC curve and its AUC value of 0.91 indicate a strong performance in distinguishing between the two income classes.
- Confusion matrix: The majority of mistakes made by the model involve false negatives (predicting '>50K' when the actual income is '<=50K'). Understanding the specific context and consequences of these mistakes is essential to evaluate the model's suitability for the intended application.
- Feature importance: Marital status is the most important feature in predicting income levels, followed by education level and capital gain.
- Potential improvements: The model's performance may be further improved by collecting more data, fine-tuning model parameters, or trying alternative machine learning algorithms.
- Model validation: Periodically validating the model using updated data will help ensure its continued accuracy and relevance in addressing the problem.

OUT-OF-SAMPLE PREDICTIONS

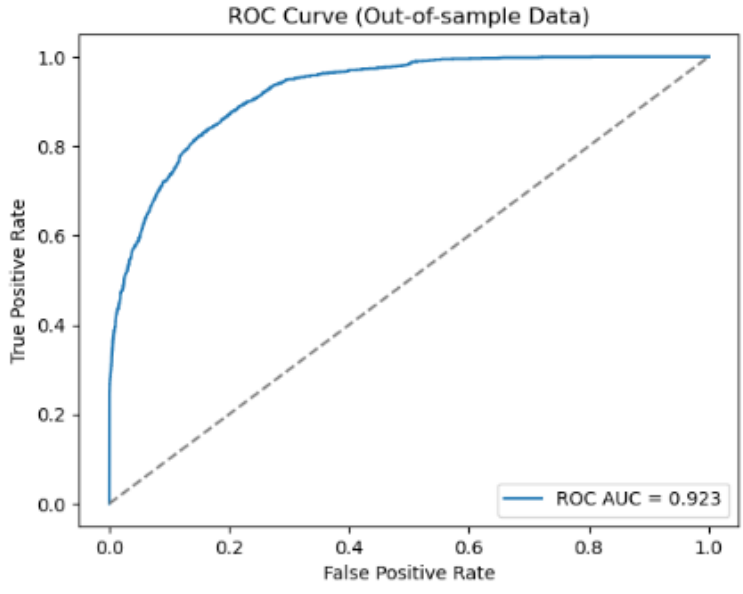
Summary of Results

```
Out-of-sample accuracy: 0.865
```



	precision	recall	f1-score	support
0	0.89	0.94	0.91	4640
1	0.78	0.63	0.70	1511
accuracy			0.87	6151
macro avg	0.83	0.79	0.81	6151
weighted avg	0.86	0.87	0.86	6151

- Model accuracy: 0.865
- Confusion Matrix:
  - True Negative (<=50K): 4366
  - False Negative: 555
  - True Positive (>50K): 956
  - False Positive: 274



- ROC curve (area = 0.923)

OUT-OF-SAMPLE PREDICTIONS

Assessment of Results

- The model achieved an accuracy of 0.865, which is comparable to the accuracy on the test data (0.857). This indicates that the model is generalizing well to unseen data.

- The confusion matrix shows the following results:
  - True Negative (<=50K): 4366
  - False Negative: 555
  - True Positive (>50K): 956
  - False Positive: 274
  - This distribution of predictions is similar to that of the test data, further supporting the model's consistent performance.

- The classification report presents the following metrics:
  - Precision: 0.78 for class 1 (>50K) and 0.89 for class 0 (<=50K)
  - Recall: 0.63 for class 1 (>50K) and 0.94 for class 0 (<=50K)
  - F1-score: 0.70 for class 1 (>50K) and 0.91 for class 0 (<=50K)
  - These metrics indicate that the model performs better at identifying individuals with incomes <=50K, but it still maintains a reasonable performance for the >50K class.

- ROC curve area 0.923, indicating strong model performance with a rapidly increasing true positive rate compared to false positive rate.

- In conclusion, the model demonstrates a consistent and strong performance on out-of-sample data, maintaining similar accuracy, confusion matrix distribution, and classification metrics as observed with the test data. This suggests that the model is generalizing well and should be effective in predicting income levels in practice.

CONCLUDING REMARKS

Summary of the analytics process

- Performed exploratory data analysis to understand the dataset and identify trends.
- Preprocessed the data, including handling missing values, encoding categorical variables, and splitting the dataset into training and test sets.
- Experimented with multiple feature selection techniques, scaling methods, and classifiers.
- Employed a cross-validated grid search to find the best combination of feature selection, scaling, and classifier.
- Evaluated the best model on test data and out-of-sample data, analyzing performance using accuracy, confusion matrix, classification report, and ROC curve.

Major findings

- Performed exploratory data analysis to understand the dataset and identify trends.
- Preprocessed the data, including handling missing values, encoding categorical variables, and splitting the dataset into training and test sets.
- Experimented with multiple feature selection techniques, scaling methods, and classifiers.
- Employed a cross-validated grid search to find the best combination of feature selection, scaling, and classifier.
- Evaluated the best model on test data and out-of-sample data, analyzing performance using accuracy, confusion matrix, classification report, and ROC curve.

CONCLUDING REMARKS

Key business (managerial) implications

- The developed model can be used to reliably predict an individual's income level, providing valuable insights for various applications, such as marketing, financial aid, or social benefits targeting.

- Identified features (e.g., marital status, education, and capital gains) can guide decision-makers in understanding the factors driving income and shaping targeted strategies.

- False positives and false negatives should be considered when evaluating the model's applicability to a specific use case, as the costs and consequences may vary.

- Model performance can be further improved by gathering more data or refining the feature set, depending on the specific requirements of a given application.

ACKNOWLEDGEMENTS

CSIS 3290 Fundamentals of Machine Learning In Data Science - Winter 2023

We would like to express our sincere gratitude to Professor Ivan Wong (Wong Tak-Lam) for his invaluable guidance, support, and mentorship throughout this study. His expertise and insights have greatly contributed to the successful completion of this report.

Presentation Link: https://youtu.be/Kj2i5I2stOU