Income Predictor : A Comprehensive Approach to Income Predicting(based on census data) Using SVM and Decision Tree Algorithms

Perera K.R.D.
EG/2020/4113
Faculty of Engineering University of Ruhuna, Sri Lanka
Hapugala, Galle
perera_krd_e22@engug.ruh.ac.lk

Suraweera S.A.Y.A.
EG/2020/4225
Faculty of Engineering University of Ruhuna, Sri Lanka
Hapugala, Galle
suraweera_saya_e22@engug.ruh.ac.lk

***Abstract*— This project explores the application of machine learning techniques to predict whether an individual's income exceeds $50,000 based on demographic and socioeconomic features from the adult census dataset. There are fourteen features in the dataset: age, race, gender, marital status, education level, and work class, and others. For predictive modeling, two effective algorithms are used: Decision Trees and Support Vector Machines (SVM). The study not only focuses on model accuracy but also addresses ethical considerations associated with predictive models, emphasizing fairness and transparency in the context of income prediction. This research adds significant insights to the field of machine learning used to macroeconomic predictions by exploring the complexities of these algorithms and their effects on society. The practical implications of the findings for understanding and explaining income inequality will benefit stakeholders and decision-makers across a range of disciplines.**

I. INTRODUCTION

In the intricate realms of financial planning and market dynamics, predicting income levels shares a common ground with forecasting used automobile prices. This project endeavors to leverage the lessons learned from automobile pricing models, applying them to the challenge of accurately predicting annual income levels exceeding $50,000. Using the Adult census dataset's diverse demographic and socioeconomic attributes, we employ Support Vector Machines (SVM) and Decision Tree algorithms for this binary classification task. Our goal is to provide a reliable and transparent income prediction system, akin to the quest for precise pricing predictions in the ever-evolving used car market. By drawing parallels between these two domains, we seek to contribute meaningful insights, empowering individuals and policymakers alike in navigating the complexities of financial planning.

II. METHODOLOGY

A. Data

The dataset for this project was sourced from the UC IRVINE website, encompassing approximately 33,000 data entries with 14 key features. These features, ranging from demographic details such as age, race, and education to socio-economic indicators like occupation and marital status, were curated from the Adult census dataset. "Income," the target variable, indicates whether or not a person makes more than $50,000 annually. The comprehensive dataset offers a solid basis for the training and assessment of the Decision Tree and Support Vector Machine (SVM) algorithms used in our predictive modeling. The features selected are in line with the various characteristics that impact income levels, which allows for a thorough investigation of trends and insights in the field of income prediction.

B. Pre-Proccesing

The first step in the income prediction project is to load and examine the "Adult Census dataset" that is available on the UC IRVINE website. The dataset is converted into a Pandas DataFrame using Pandas tools, allowing for a thorough examination of its structure. It's important to address missing entries because in the dataset in "workClass" column there were 1836 null values, in "NativeCountry" 583 and "Occupation" 1843 null values.

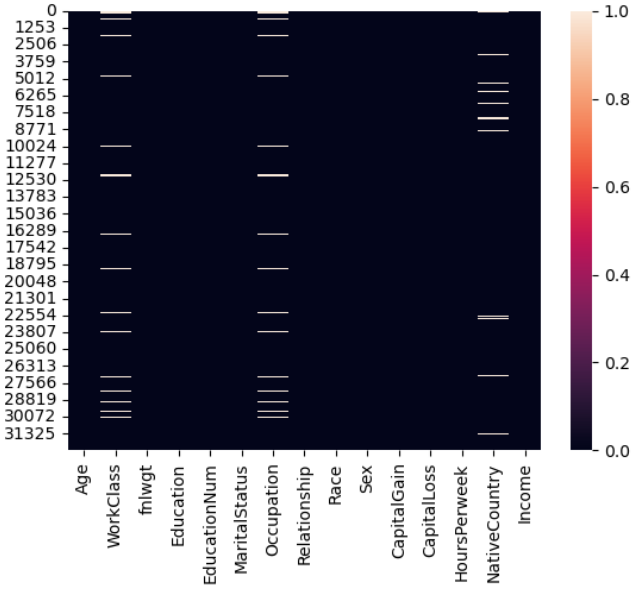

Figure 1:Heat map for the null values

For Null values for "workclass" and "Nativecountry" features, null values are replaced by mode value and the null values of "Occupation" are deleted. There were 24 duplicate rows and those are deleted. There were two duplicate rows "EducationNum" and "Education" and "EducationNum" column was deleted.

To improve predictive power, feature engineering is used to combine distinct values in every category column according to the percentage of earnings that is more than $50,000. Then we checked the outliers of numerical columns and didn't do the outlier treatment because SVM and decision tree algorithms are not much sensitive for outliers. Then we used PCA algorithm for dimentianility reduction and total number of features is reduced to 1 and I removed PCA.Categorical data is converted into numerical format using one-hot encoding, and numerical features are scaled using StandardScaler.

After the data is properly divided into features and target variables and outliers are treated, the thorough pre-processing ensures an accurate and improved dataset for the next stage of model construction. After undergoing these thorough pre-processing procedures, the final dataset provides a strong basis for the construction and assessment of the model.

C. Algorithms

Support Vector Machines (SVM):
Support Vector Machines are powerful classifiers capable of handling both linear and non-linear relationships in datasets. SVM is used in our income prediction model to find the best possible hyperplane that divides people who earn more than $50,000 from those who don't. The process of maximizing the margin between the two classes results in this binary classification. SVM works well with high-dimensional data and is capable at identifying complex patterns. By optimizing parameters like kernel type and regularization, GridSearchCV helps the model be tuned and ensures the most accurate split of the income groups.

Decision trees:
On the other hand, Decision Tree is selected for its ability to capture complex relationships in the data. In order to estimate income levels, our project uses the Decision Tree algorithm to analyze the Adult Census information and identify trends in socioeconomic and demographic factors. This approach works well at capturing complex relationships in the data, which makes it useful in situations when linear models might not be sufficient. Using GridSearchCV, we adjust hyperparameters such as maximum depth and minimum samples split to optimize the tree's structure and forecast whether a person makes more than $50,000 annually.

D. Implementation

This income prediction project's implementation phase follows to a defined methodology that includes critical steps such as data splitting, preprocessing, model training, hyperparameter tuning, and thorough evaluation. These processes are carried out with simplicity due to the flexibility of the sci-kit-learn library, a popular Python machine learning toolkit.

Data Splitting:
To assess model performance on untested data and prevent overfitting, the dataset is divided into training and testing sets at a 0.8 to 0.2 ratio. This intentional separation ensures an accurate inspection of algorithmic prediction performance.

Preprocessing:
A preprocessing pipeline is established to effectively handle numerical and categorical features. To identify any non-linear connections, standard scaling is applied to numerical characteristics. One-hot encoding is used to encode categorical characteristics, guaranteeing compatibility with the Decision Tree and SVM algorithms that are being used.

Model Training:
The Decision Tree and SVM models are trained using the preprocessed training data. Using the scikit-learn package, the training process may be made more efficient and simple by making advantage of the algorithms' basic capabilities.

Hyperparameter Tuning:
GridSearchCV from scikit-learn facilitates systematic hyperparameter adjustment to optimize model performance. Parameters such as “kernel”, ”gamma” and “c” are used for SVM, and “max_depth”, “min_samples_split”, “min_sample_leaf”, “max_features”, “criterion” are leaf for Decision Trees, to go through a thorough adjusting process to determine the best configurations.

Evaluation:
Model performance is assessed using confution matrix, providing a numerical representation of the variance in the target variable explained by the models. The scikit-learn module enables the calculation of these measures, guaranteeing an accurate and unbiased assessment.

III. RESULTS

In this section, we present a detailed analysis of the results obtained from both SVM and Decision Tree algorithms. We compare the initial and fine-tuned performance metrics, including accuracy, precision, recall, and F1-score and cross-validation scores. Furthermore, the impact of the additional dataset on training and testing accuracy is explored, providing insights into the effectiveness of data augmentation.

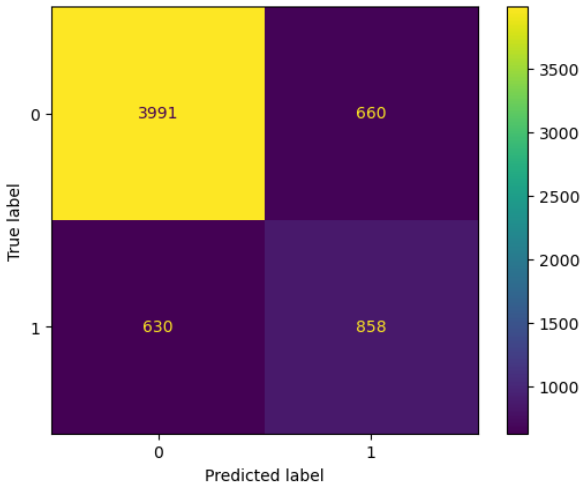

Figure 2: Confusion matrix for Decision Tree.

	precision	recall	f1-score	support
0	0.86	0.92	0.89	4651
1	0.68	0.53	0.60	1488
accuracy			0.83	6139
macro avg	0.77	0.73	0.74	6139
weighted avg	0.82	0.83	0.82	6139

Figure 3: Classification report for Decision Tree

The accuracy of the Decision Tree model can be shown from the confusion matrix. Further, the accuracy, precision and recall can be easily taken from the classification report.

According to the final result of the decision tree algorithm, income levels over $50,000 reveals that, among instances predicted as having higher income, the model's precision stands at 68%, signifying the accuracy of positive predictions. The recall, measuring the model's ability to capture all actual instances with income over $50,000, is 53%. The F1-score, a balanced assessment combining precision and recall, is calculated at 0.60. Overall accuracy, reflecting correctness across both classes, is 83%.

The overall test accuracy of this decision tree model is 0.82.

Same as the Decision Tree the performance of the SVM can be taken from the confusion matrix and the classification report.

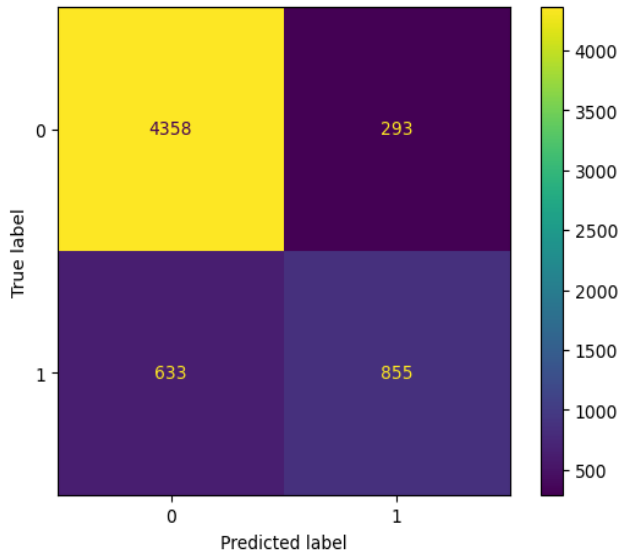

Figure4: Confusion matrix for SVM

	precision	recall	f1-score	support
0	0.87	0.94	0.90	4651
1	0.74	0.57	0.65	1488
accuracy			0.85	6139
macro avg	0.81	0.76	0.78	6139
weighted avg	0.84	0.85	0.84	6139

Figure5: Classification report for SVM

According to the final result of the svm algorithm, the classification report evaluates the model's effectiveness in predicting individual's income. For instances predicted as having higher income, the precision is 74%, signifying a reasonably accurate positive prediction rate. The recall, measuring the model's ability to correctly identify actual instances with income over $50,000, is at 57%, indicating a moderate capturing of positive instances. The F1-score, a balanced metric combining precision and recall, stands at 0.65, suggesting a reasonable compromise between the two.

The overall test accuracy of this decision tree model is 0.85.

In conclusion, according to the performance measures, SVM has performed well than the Decision Tree model. With the given results, based on the features in the adult census dataset can do the accurate predictions.

IV. Discussion

In our project about predicting income using Adult Census data, we used two types of models: Support Vector Machines (SVM) and Decision Trees. We chose these models because each has its own strengths and weaknesses. We carefully considered how they work to predict people's income.

Support Vector Machines (SVM) are good at understanding linear trends, which are straightforward patterns in the data. They help us figure out relationships in the information we have. On the other hand, Decision Trees are excellent at spotting more complex, non-linear patterns. This makes them useful for handling the diverse nature of predicting income.

To make reliable predictions, we needed to balance reliability and flexibility in these models. SVMs provide a reliable foundation by understanding linear relationships well. They help us make sense of simple trends in the data. Decision Trees, on the other hand, offer flexibility by being great at understanding non-linear patterns, handling the complexity of real-world income data.

However, this flexibility also brings challenges, especially the risk of overfitting. Overfitting happens when a model gets too focused on the details in the data it was trained on, making it less effective with new data. Striking the right balance

between reliability and flexibility is crucial, like finding an artful mix to make sure our predictions are trustworthy and adaptable to different real-world income situations.

In summary, we chose SVMs and Decision Trees because they complement each other's strengths in predicting income levels. As we work through the complexities of this task, we're aware of the need for a balanced approach. We keep refining our models to find the right mix of reliability and flexibility for accurate predictions in various income scenarios.

Ethics is a key element in the development and implementation of an Income Predictor based on census data. Given that this project involves sensitive information related to individuals and their financial status, it is imperative to adhere to ethical considerations throughout the entire process. The ethical principles outlined below are essential for guiding the development of the Income Predictor and ensuring its responsible use.

Fairness:
In ensuring fair model predictions, we focus on evaluating and mitigating biases. Before training the models, we thoroughly examine the dataset for any inherent biases and take steps to minimize their impact. We use fairness metrics, such as demographic parity and equalized odds, to assess and improve the fairness of the models. By actively addressing fairness concerns, our project aims to avoid perpetuating or worsening existing inequalities.

Transparency:
Model transparency is crucial for building trust and understanding how predictions are made. We use decision trees to create a more interpretable model, allowing the decision-making process to be visualized and understood. Additionally, we thoroughly document the hyperparameter tuning process to provide transparency in the choices made during model development. This transparency makes the model more accessible and understandable for stakeholders, promoting responsible and informed use.

Privacy Protection:
We prioritize protecting individuals' privacy, taking steps to ensure that prediction models do not disclose sensitive personal information beyond what is necessary for the intended purpose. We employ anonymization and data aggregation techniques to reduce the risk of re-identification of individuals in the dataset.

Informed Consent:
Individuals whose data is used for model training and testing are informed about the study's purpose and how their data will be used. Obtaining explicit consent, when feasible, ensures individuals are aware of and agree to the use of their data for predictive modeling.

Data Security:
We implement robust measures to safeguard the dataset from unauthorized access, data breaches, or malicious use. This includes secure storage, data encryption, and access controls. Transparent communication about the security measures in place instills confidence among users and stakeholders.

Long-Term Societal Impact:
Consideration is given to the potential long-term societal impact of predictive models. Understanding how deploying these models may affect social structures, economic dynamics, and individual opportunities is crucial. Ongoing research and collaboration with social scientists, ethicists, and policymakers contribute to a deeper understanding of the broader societal implications.

In conclusion, this research carefully used two types of models, Support Vector Machines (SVM) and Decision Trees, to estimate income levels using the Adult Census data. Combining these models worked well, giving us accurate results. The research involved looking at the data, preparing it for analysis, using computer programs (algorithms), and adjusting the models to make them better.

We made sure to test the models thoroughly by using a method called cross-validation, which helps ensure they work well in different situations. The study highlights the importance of using data responsibly, especially when dealing with private information like income. It also shares practical ways to find subtle patterns in income data, which can be helpful for economic predictions. It provides useful information for decision-makers, not just in the academic world but also for those dealing with the challenges of economic predictions in real life. Essentially, this research guides decision-makers on how to use machine learning responsibly in situations where privacy and ethics matter.

.

V. REFFERENCE

[1] "UC Irvine," [Online]. Available: https://archive.ics.uci.edu/dataset/2/adult . [Accessed 15 11 2023].

[2] "scikit learn," [Online]. Available: https://scikit-learn.org/stable/modules/svm.html . [Accessed 12 12 2023].

[3] "scikit learn," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html. [Accessed 10 1 2024].

[4] "ChatGpt," [Online]. Available: https://chat.openai.com/. [Accessed 2024].