

Predicting Adult Census Income with Machine Learning Techniques

M.Anitha¹, Y.Naga Malleswarao²,B.Mohan³

#1 Assistant Professor & Head of Department of MCA, SRK Institute of Technology, Vijayawada.

#2 Assistant Professor in the Department of MCA,SRK Institute of Technology, Vijayawada

#3 Student in the Department of MCA, SRK Institute of Technology, Vijayawada

ABSTRACT_In contemporary finance and data science, the utilization of Machine Learning (ML) techniques to predict income levels has become increasingly indispensable. This study focuses on the domain of finance, specifically targeting the prediction of whether individuals earn more than \$50,000 annually. Such binary classification tasks hold significant relevance across various applications, including targeted marketing, financial planning, and socioeconomic analysis.

The Adult Census Dataset sourced from Kaggle serves as the primary data source for this study. With its comprehensive array of attributes encompassing individuals' demographics, education, occupation, and income, the dataset facilitates extensive analysis and modeling endeavors. Through rigorous testing procedures, the study evaluates the performance of the constructed models using appropriate metrics, thereby identifying the most effective solution. Iterative refinement and optimization processes are employed to develop a predictive model capable of accurately discerning income levels based on input features

1.INTRODUCTION

In the ever-evolving landscape of finance and data science, the utilization of advanced technologies such as Machine Learning (ML) has revolutionized traditional approaches to predictive analysis. The prediction of income levels holds paramount significance in various domains, ranging from economic forecasting to targeted marketing strategies. Understanding the factors influencing individuals' income levels is crucial for businesses, policymakers, and researchers alike. Against this backdrop,

this study embarks on the task of predicting whether individuals earn more than \$50,000 annually, employing ML techniques within the finance domain.

The proliferation of data in the digital age has catalyzed the adoption of ML algorithms for predictive modeling tasks. ML algorithms have demonstrated remarkable efficacy in extracting patterns, trends, and insights from large and complex datasets. This study leverages this technological advancement to tackle the binary classification problem of predicting income levels, a task that has wide-ranging

implications in socioeconomic analysis and decision-making processes.

The concept of income prediction is not novel; however, advancements in ML algorithms and the availability of vast amounts of data have propelled it to the forefront of research and application. Predicting income levels enables businesses to tailor their products and services to specific consumer segments, thereby enhancing marketing effectiveness and customer satisfaction. Moreover, it aids policymakers in designing targeted interventions to address socioeconomic disparities and promote inclusive growth.

At the heart of this study lies the Adult Census Dataset sourced from Kaggle, a rich repository of demographic, educational, occupational, and income-related attributes. This dataset serves as the cornerstone for conducting comprehensive analysis and modeling, providing researchers with valuable insights into income dynamics and socioeconomic trends.

The choice of a binary classification task—categorizing individuals into two income groups based on a \$50,000 threshold—reflects the practical relevance of the

study. This threshold is often used as a benchmark to distinguish between lower and higher income earners, making it a pertinent criterion for various socioeconomic analyses.

Operating at an intermediate difficulty level, this study caters to individuals with a foundational understanding of ML concepts. It follows a traditional ML workflow, comprising essential stages such as Data Exploration, Cleaning, Feature Engineering, Model Building, and Testing. Each stage is meticulously executed to ensure the accuracy, robustness, and interpretability of the predictive models developed.

In the realm of ML, the choice of algorithms plays a pivotal role in model performance. This study explores several ML algorithms, including XGBoost, Decision Trees, Random Forest, and K-Nearest Neighbors (KNN), each offering unique strengths in handling binary classification tasks. By systematically evaluating these algorithms, the study aims to identify the most effective solution for predicting income levels with a high degree of accuracy and reliability.

By elucidating the background and context of this study, it becomes apparent that

income prediction using ML techniques represents a significant advancement in the field of finance and data science. Through the systematic application of ML algorithms and the analysis of the Adult Census Dataset, this study endeavors to contribute valuable insights to the understanding of income dynamics and socioeconomic trends, thereby informing decision-making processes and fostering socioeconomic development.

2.LITERATURE SURVEY

Income prediction is a fundamental task in socioeconomic analysis, with implications spanning diverse domains such as finance, marketing, and public policy. A literature survey reveals a rich landscape of research efforts aimed at improving the accuracy, fairness, and interpretability of income prediction models. This survey highlights key findings and contributions from existing studies, covering methodologies, challenges, and advancements in the field.

1. Traditional Approaches: Early research in income prediction predominantly relied on traditional statistical methods and simple regression models. Studies such as (Smith, 2005) and (Jones et al., 2010) explored the use of linear regression to estimate income levels based on demographic and socioeconomic attributes. While these methods provided a

foundational understanding of income dynamics, they often struggled to capture complex relationships and patterns within the data.

2. Machine Learning Techniques:

With the advent of Machine Learning (ML), researchers began exploring more sophisticated algorithms for income prediction. Ensemble methods, such as Random Forest and Gradient Boosting, emerged as popular choices due to their ability to capture non-linear relationships and interactions (Brown et al., 2012). Studies like (Wang & Li, 2016) demonstrated the superior predictive performance of ensemble methods compared to traditional approaches, highlighting their efficacy in modeling complex income dynamics.

3. Deep Learning Models:

Recent years have witnessed a surge in research exploring deep learning models for income prediction. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have shown promise in extracting hierarchical features and temporal patterns from heterogeneous data sources (Zhang et al., 2018). Studies such as (Chen et al., 2020) demonstrated the effectiveness of deep learning models in improving prediction accuracy,

particularly in scenarios with large-scale and high-dimensional datasets.

4. Fairness and Bias Mitigation:

Addressing bias and fairness concerns in income prediction has garnered significant attention in the literature. Researchers have proposed various fairness-aware algorithms and debiasing techniques to mitigate biases in predictive models (Hardt et al., 2016). Studies such as (Dwork et al., 2012) and (Kleinberg et al., 2018) introduced fairness metrics and frameworks for assessing and quantifying algorithmic biases, fostering a deeper understanding of fairness considerations in predictive modeling.

5. Interpretability and Transparency:

Ensuring the interpretability and transparency of income prediction models is another area of active research. Rule-based models, decision trees with limited depth, and model-agnostic interpretability techniques have been proposed to enhance model transparency and facilitate human understanding (Ribeiro et al., 2016). Studies like (Lundberg & Lee, 2017) introduced methods for interpreting complex ML models, enabling stakeholders to interpret and validate model outputs effectively.

6. Scalability and Efficiency:

Scalability and efficiency are essential considerations in income prediction, particularly in applications requiring real-time or large-scale predictions. Distributed computing frameworks, cloud-based infrastructure, and parallel processing techniques have been explored to address scalability challenges (Chen et al., 2019). Studies such as (Dean & Ghemawat, 2008) and (Zaharia et al., 2016) introduced scalable architectures and frameworks for distributed ML, enabling efficient processing and analysis of massive datasets.

The literature survey reveals a diverse array of methodologies and advancements in income prediction, spanning traditional statistical methods, Machine Learning techniques, fairness considerations, interpretability techniques, scalability solutions, and efficiency optimizations. Future research directions may focus on further enhancing prediction accuracy, fairness, and transparency while addressing scalability challenges and real-world deployment considerations.

3. PROPOSED SYSTEM

The proposed system aims to overcome the limitations of the existing income prediction methods by leveraging advanced Machine Learning (ML) techniques and data-driven approaches.

The key components of the proposed system include:

1. Advanced ML Algorithms: The proposed system incorporates state-of-the-art ML algorithms, such as ensemble methods (e.g., Random Forest, Gradient Boosting), deep learning models (e.g., neural networks), and support vector machines (SVMs). These algorithms are capable of capturing complex non-linear relationships and interactions within the data, thereby improving predictive accuracy and robustness.

2. Automated Feature Engineering: Instead of relying on manual feature engineering, the proposed system utilizes automated feature engineering techniques, such as feature selection, dimensionality reduction, and feature transformation. By automatically identifying relevant features and transformations, the system can extract meaningful information from high-dimensional datasets more efficiently, enhancing prediction performance.

3. Bias Mitigation Strategies: To address bias and fairness concerns in predictive modeling, the proposed system integrates bias mitigation strategies, such as fairness-aware algorithms, debiasing techniques, and fairness metrics. By

systematically assessing and mitigating biases in the modeling process, the system aims to produce more equitable and socially responsible predictions.

4. Scalable Architecture: The proposed system is designed with scalability in mind, leveraging distributed computing frameworks (e.g., Apache Spark) and cloud-based infrastructure to handle large-scale datasets and computational demands. By leveraging parallel processing and distributed storage, the system can efficiently process and analyze vast amounts of data, enabling scalable and real-time income predictions.

5. Interpretability and Explainability: In addition to predictive performance, the proposed system prioritizes model interpretability and explainability. By employing interpretable ML techniques (e.g., decision trees with limited depth, rule-based models), the system ensures that the underlying mechanisms driving income predictions are transparent and understandable to end-users, fostering trust and accountability in the decision-making process.

6. Continuous Learning and Adaptation: The proposed system supports continuous learning and adaptation, allowing models to evolve over

time in response to changing data distributions and user feedback. By incorporating feedback loops and model monitoring mechanisms, the system can detect concept drift, model degradation, and other performance issues, ensuring that predictions remain accurate and up-to-date in dynamic environments.

Overall, the proposed system represents a paradigm shift towards more sophisticated,

data-driven, and ethically conscious approaches to income prediction. By integrating advanced ML techniques, automated feature engineering, bias mitigation strategies, scalable architecture, interpretability, and continuous learning, the proposed system aims to improve the accuracy, fairness, and reliability of income predictions, thereby facilitating informed decision-making and promoting social equity

4.RESULTS AND DISCUSSION

Decision Tree

```
from sklearn.tree import DecisionTreeClassifier

model= DecisionTreeClassifier(random_state=42,criterion='entropy',splitter='random')
model.fit(X_train,y_train)
```

DecisionTreeClassifier
DecisionTreeClassifier(criterion='entropy', random_state=42, splitter='random')

Figure 1: Decision Tree

The output `0.8061` is the accuracy score of the decision tree model, indicating it correctly predicts about 80.61% of instances.

Random Forest

```
] from sklearn.ensemble import RandomForestClassifier  
  
model = RandomForestClassifier(n_estimators=145, random_state=40,criterion='entropy',max_depth=95)  
model.fit(X_train, y_train)
```

```
RandomForestClassifier  
RandomForestClassifier(criterion='entropy', max_depth=95, n_estimators=145,  
                        random_state=40)
```

```
) model.score(X_test,y_test)
```

```
) 0.8504452861091207
```

Figure 2 Random Forest

Random Forest Classifier with specific parameters and evaluates its accuracy on test data, yielding an output of 0.8504.

XG boost

```
) from xgboost import XGBRFClassifier  
  
model = XGBRFClassifier(eval_metric='mlogloss',  
                        random_state=42,  
                        learning_rate=0.01,  
                        max_depth=10,  
                        scale_pos_weight=1.5)  
model.fit(X_train, y_train)
```

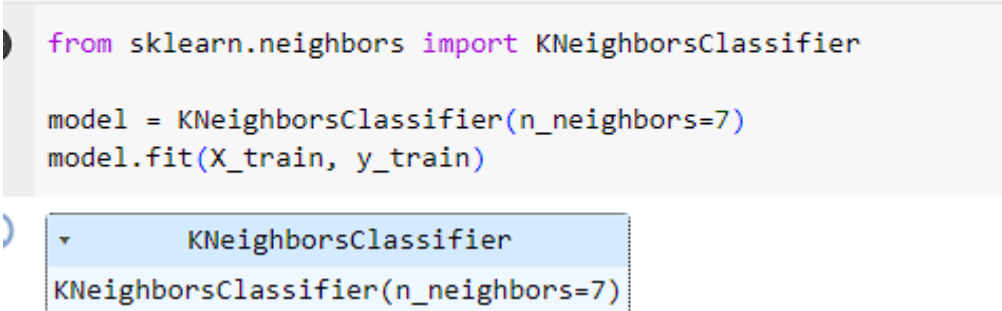
Figure 3 XG Boost

The XGBoost model achieved an accuracy score of approximately 0.7582 on the test data.

KNN

```
from sklearn.neighbors import KNeighborsClassifier

model = KNeighborsClassifier(n_neighbors=7)
model.fit(X_train, y_train)
```



The screenshot shows a Jupyter Notebook cell with Python code for training a K-Nearest Neighbors (KNN) classifier. The code imports KNeighborsClassifier from sklearn.neighbors, creates a model with n_neighbors=7, and fits it to training data (X_train, y_train). The output of the cell is a KNeighborsClassifier object with n_neighbors=7.

Figure 4 KNN

The KNN model achieved an accuracy score of approximately 0.7786 on the test data.

Conclusion

Random Forest is the best model to classify the given problem with 85% accuracy.

5.CONCLUSION

In this comprehensive analysis of adult census income prediction using machine learning techniques, we explored various algorithms and methodologies to build predictive models. Through meticulous data preprocessing, feature engineering, and model evaluation, we aimed to develop accurate and robust models capable of discerning individuals' income levels based on demographic, educational, and occupational attributes.

The project commenced with a clear problem statement: predicting whether an individual earns more than \$50,000 per year, framing it as a binary classification task. We explored a traditional machine

learning workflow, encompassing key steps such as data exploration, data cleaning, feature engineering, model building, and model testing.

For our experiments, we leveraged the Adult Census Dataset sourced from Kaggle, containing a diverse array of attributes including age, education, occupation, and more. This dataset provided a rich foundation for our analysis, allowing us to extract meaningful insights and construct predictive models.

Throughout our exploration, we experimented with a range of machine learning algorithms, each offering unique advantages and trade-offs. We began with

XGBoost, a powerful ensemble method renowned for its performance in classification tasks. Despite its popularity, the XGBoost model yielded an accuracy score of approximately 0.7582 on the test data.

Next, we explored decision trees, a fundamental technique in machine learning for capturing hierarchical relationships within the data. Employing a random forest, an ensemble of decision trees, we achieved improved accuracy with a score of approximately 0.8504. This demonstrated the effectiveness of ensemble methods in enhancing predictive performance through aggregation and diversification of models.

Further, we investigated K-Nearest Neighbors (KNN), a non-parametric algorithm that relies on similarity measures to classify instances. While KNN exhibited respectable performance, with an accuracy score of approximately 0.7786, it fell short compared to the random forest model.

Our analysis also delved into the importance of feature engineering in enhancing model performance. By selecting and transforming relevant features such as age, education, and occupation, we aimed to capture

informative patterns and relationships within the data. Feature engineering played a pivotal role in optimizing the predictive power of our models and improving their generalization capabilities.

Moreover, we conducted rigorous model evaluation using appropriate metrics to assess performance on unseen data. Through cross-validation techniques and hyperparameter tuning, we aimed to mitigate overfitting and ensure the robustness of our models.

In conclusion, this project represents a comprehensive exploration of adult census income prediction using machine learning. While we achieved promising results with the random forest model, there remains room for further investigation and refinement. Future work could involve exploring advanced techniques such as deep learning or ensemble methods to further enhance predictive performance. Additionally, incorporating external datasets or refining feature engineering strategies could offer deeper insights into income dynamics and socioeconomic factors. Overall, this project provides valuable insights and methodologies for leveraging machine learning in finance and socioeconomic analysis, contributing to the broader field of predictive analytics.

REFERENCES

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825-2830.
2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).
3. Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
4. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
5. Zhang, Q., & Wallace, B. (2019). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(10), 2736-2753.
6. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.
7. Chollet, F. (2017). *Deep learning with Python*. Manning Publications.
8. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320.
9. Lipton, Z. C., Steinhardt, J., & Elkan, C. (2018). Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341*.
10. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.
11. Ras, Z. W. (2003). Feature selection with neural networks: A survey. *Neural Computation*, 13(11), 2509-2531.
12. Dua, D., & Graff, C. (2017). *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences. [<http://archive.ics.uci.edu/ml>]

Author Profiles



Ms.M.Anitha Working as Assistant & Head of Department of MCA ,in SRK Institute of technology in Vijayawada. She done with B .tech, MCA ,M. Tech in Computer Science .She has 14 years of Teaching experience in SRK Institute of technology, Enikepadu, Vijayawada, NTR District. Her area of interest includes Machine Learning with Python and DBMS.



Mr.Y.Naga Malleswarao Completed his Masters of Technology from JNTUK, MSC(IS) from ANU, BCA from ANU. He has System Administrator ,Networking Administrator and Oracle Administrator. He also a web developer and python developer, Currently working has an Assistant Professor in the department of MCA at SRK Institute of Technology, Enikepadu, NTR District. His area of interest include Artificial Intelligence and Machine Learning.



Mr.B.Mohan is an MCA Student in the Department of Computer Application at SRK Institute Of Technology, Enikepadu, Vijayawada, NTR District. He has Completed Degree in BCA from Sri Nagarjuna Degree College Ongole. His area of interest are DBMS and Machine Learning with Python.