Artificial Intelligence in Applied Sciences
Cutting-Edge Scientific Solutions

e-ISSN: 3108-4060
RESEARCH ARTICLE
Vol.1 / No.1 / 2025 / pp.7-12
https://doi.org/10.69882/adba.ai.2025072

Income Level Estimation with Light-GBM: Understanding Model Decisions with Explainable AI Techniques Shap and Lime

Cem Özkurt [*,α,1], **Fatmir Garri** [α,2], **Bilal Emre Yahyaoğlu** [α,3], **Onur Ağca** [α,4], **Necip Furkan Bildiren** [α,5] **and Sergen Kaynak** [α,6]

[*] AI and Data Science Research and Application Center, Sakarya University of Applied Science, Türkiye, [α] Department of Computer Engineering, Sakarya University of Applied Sciences, Türkiye.

ABSTRACT This study examines the use of machine learning and artificial intelligence algorithms to predict individuals' annual incomes. In analyses conducted using the Python programming language, the best performance was achieved in models utilizing the "Synthetic Minority Over-sampling Technique (SMOTE)" for imbalanced data sets, with an accuracy of 87.45%, precision of 85.74%, recall of 89.31%, and an F1 score of 87.30, using the "Light Gradient Boosting Machines" algorithm. Additionally, the impact of parameters and variables on income prediction was examined using interpretable artificial intelligence algorithms. The results of the study emphasize the importance of employing effective methods and explaining machine learning model predictions, as well as addressing imbalanced data sets.

KEYWORDS
Revenue prediction
Machine learning
Explainable artificial intelligence
SMOTE

INTRODUCTION

The estimation of individuals' incomes is crucial for financial planning and resource management. This study aims to explore practical applications of machine learning and artificial intelligence algorithms to predict individuals' annual incomes. The research seeks to develop income prediction models and achieve more accurate estimations. Several studies focus on estimating individual income levels using various approaches. One study employed a machine learning approach to predict individual income and highlighted the issue of individuals misreporting their earnings (Matkowski 2021). Another study explored individual-level income prediction using Facebook profiles, examining the density distributions of annual income and comparing them with U.S. Census data (Matz *et al.* 2019). Additionally, the UCI Adult Dataset, a common resource for predicting annual income levels in the U.S., has been used to classify whether a person's income exceeds a certain threshold. This dataset has been applied to predict whether an individual's annual income surpasses $50,000 based on demographic data (Becker and Kohavi 2023).

An analysis was conducted on subjective expectations about future income changes using household panel data. This study found that income changes strongly depended on past changes. It also observed that expected income changes were significantly influenced by factors such as employment status, family structure, permanent income, and past expectations. The study concluded that expectations were not rational, particularly noting that households with decreasing past incomes underestimated future income growth (Das and Van Soest 1999). Another research examined weekly earnings expectations reported in subjective probabilities by participants in a national household survey during the spring of 1994. This study assessed the potential of obtaining expectations in future surveys, suggesting that such data could be more informative than typical economic expectations reports.

It also analyzed revisions in expectations and the relationship between expectations and actual earnings, providing positive findings on the validity of the data (Dominitz 1998). One study investigated the use of principal component analysis and support vector machines to create and evaluate income prediction data based on the U.S. Census Bureau's Current Population Survey. This research demonstrated the effectiveness of detailed statistical studies for rel-

Manuscript received: 25 November 2024,
Revised: 6 June 2025,
Accepted: 12 June 2025.

[1] cemozkurt@subu.edu.tr **(Corresponding author)**
[2] 23502405002@subu.edu.tr
[3] 24502405018@subu.edu.tr
[4] 24502405013@subu.edu.tr
[5] 24502405013@subu.edu.tr
[6] 24502405008@subu.edu.tr

evant feature selection and their impact on improving classification accuracy. It also emphasized that shaping computational methods around specific real datasets is a critical factor in enhancing the power of algorithms (Lazar 2004).

MATERIALS AND METHODS

Dataset

In this study, the dataset was examined across 14 parameters (variables or features) (Becker and Kohavi 2023). Initially containing 48,842 observations, the dataset was reduced to 45,223 observations by removing those with missing information. The dataset was organized to serve as a suitable input for all machine learning algorithms used in the study. As part of this organization, a proper format conversion was applied to parameters that did not have numerical values.

Table 1 Parameters and Data Types of the Dataset Used

Parameter	Data Type
Age	Integer
Work class	Categorical
Education Level	Categorical
Marital Status	Categorical
Occupation	Categorical
Relationship	Categorical
Race/Ethnicity	Categorical
Gender	Categorical
Capital Gain	Integer
Capital Loss	Integer
Weekly Work Hours	Integer
Native Country	Categorical
Income	Binary

Machine Learning

Machine learning algorithms are computational models that enable computers to make data-driven predictions. Supervised machine learning involves training algorithms using a labeled dataset. This type of learning allows the model to learn a mapping from input data to output labels, enabling it to make predictions or classifications on unseen data (Michalski *et al.* 2013). In this study, six different machine learning algorithms were employed, and the results were obtained.

Decision Trees: Decision trees are a machine learning algorithm used to solve classification and regression problems. This algorithm analyzes the features in a dataset to reach a conclusion through a series of decisions. The most significant advantage of decision trees is their ease of interpretation. The general formula for a decision tree is as shown in (1):

$$f(x) = \sum_{m=1}^{M} c_m \cdot I(x \in R_m) \tag{1}$$

In this equation, $f(x)$ represents the predicted output for the input feature vector x. M enotes the total number of nodes in the tree. R_m represents the region at node m. The indicator function $I(x \in R_m)$ takes the value 1 if x belongs to the region R_m and 0 otherwise. c_m denotes the predicted value at node, m.

Random Forest: A random forest is an ensemble algorithm that combines multiple decision trees. The main idea behind the mathematical formulation of the random forest model is to aggregate the predictions of each decision tree, either by averaging them (for regression tasks) or by voting (for classification tasks) (Erdem *et al.* 2018).

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \tag{2}$$

In this formula, $f(x)$ represents the predicted target variable. N denotes the total number of trees, and $f_i(x)$ represents the prediction of the i-th tree for the input dataset x. This formula takes the predictions of each tree and then averages these predictions or performs voting to arrive at the final prediction. This approach helps make the model more stable and generally improves performance, as the error tendency of one tree can offset the errors of other trees. Additionally, the algorithm introduces randomness in tree construction, ensuring that each tree is different from the others.

Gradient Boosting : Gradient boosting is a machine learning method often employing tree-based algorithms. Its primary goal is to create a strong predictive model by combining weak learners (usually decision trees) (Atasoy and Demiröz 2021; Friedman 2001). Let the dataset consist of points (x_i, y_i) , $i = 1, 2, 3, ..., N$. Here x_i represents the input features, and, y_i represents the target variable. If the model's initial prediction is set to zero:

$$F_0(x) = 0 \tag{3}$$

In each iteration, a new prediction model is added to minimize the error function.

$$F_m(x) = F_{m-1}(x) + \rho \cdot h_m(x) \tag{4}$$

In this equation, m represents the number of iterations, ρ represents the learning rate and $h_m(x)$ represents the newly added weak model.

Extreme Gradient Boosting (XGBoost): Extreme Gradient Boosting (XGBoost) is a machine learning algorithm and fundamentally a tree-based model. This algorithm is an ensemble model that combines a series of weakly learned Decision Trees. These trees are structured to complement each other and correct errors (Chen *et al.* 2019; Mitchell and Frank 2017). The Extreme Gradient Boosting algorithm is frequently used to solve regression and classification problems, and its mathematical function is generally as shown in (5):

$$F(x) = L(\theta) + \Omega(\theta) \tag{5}$$

In this equation, $L(\theta)$ represents the loss function, measuring how far the model's predictions deviate from the actual values. For classification problems, cross-entropy functions can be used as the loss function.

$$L(\theta) = \sum_{i=1}^{n} \left(-y_i \log(\hat{y}_i) + (1 - y_i)\log(\hat{y}_i)\right) \tag{6}$$

In this equation n represents the total number of data points, $\hat{y}_i$, represents the actual value of the i-th data point, and $\hat{y}_i$, represents the model's prediction for the i-th data point.

$$\Omega(\theta) = \gamma T + \frac{\lambda}{2} \sum_{j=1}^{T} w_j^2 \tag{7}$$

In this equation T represents the number of trees and w_j represents the node weights of the j-th tree. γ adds a regularization term to each tree and controls the addition of trees. λ controls the complexity of the tree by penalizing the square of the node weights. The last term of Equation (5), θ, represents the parameters of the model. These parameters include the decision rules at the nodes of each tree, the weights, and other features.

Adaptive Boosting (AdaBoost): Adaptive Boosting is an ensemble learning algorithm that combines multiple weak learners to create a strong learner. This algorithm weights each learner based on its misclassification rate after training it, using a weighted error function (Bulut 2016). This error represents the difference between the actual label y_i and the predicted label $(h_t(x_i))$.

$$\epsilon = \sum_{i=1}^{N} w_{(t,i)} \cdot \prod (h_t(x_i) \neq y_i) \tag{8}$$

In this equation, N represents the number of data points, $w_(t,i))$ is the weighting factor of the t-th learner, and $\Pi(\cdot)$ is the representative indicator function.

Weights are assigned to the learners using the formula in (9).

$$a_t = \frac{1}{n} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right) \tag{9}$$

In this formula, ϵ_t, represents the weighted error rate. The assigned weights depend on the performance of the learner. To update the weights, the formula in (10) is used.

$$w_{t+1,i} = \frac{w_{t,i} \cdot e^{-a_t \cdot y_t \cdot h_t(x_i)}}{Z_t} \tag{10}$$

In this equation, Z_t is the normalization factor that ensures the sum of the weights equals 1. With the contributions of all learners, a strong learner is created using the formula below. In this way, the Adaptive Boosting algorithm combines a series of weak learners to form a strong learner.

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} a_t \cdot h_t(x) \right) \tag{11}$$

Light Gradient Boosting Machines (LightGBM): Light Gradient Boosting Machines (LightGBM) is an implementation of the Gradient Boosting framework, a machine learning framework. Therefore, the mathematical formula of the LightGBM algorithm generally resembles the formula of the Gradient Boosting algorithm. LightGBM stands out with features such as histogram-based learning and scaled gradient descent. Its basic mathematical formula is as shown in (12).

$$F_m = F_{m-1}(x) + \eta \cdot h_m(x) \tag{12}$$

In this equation, $F_m(x)$, represents the total prediction after adding the m-th . $F_{(m-1)}(x)$, is the prediction after m-1 trees have been added. η represents the learning rate and $h_m(x)$, is the contribution of the m-th tree. Light Gradient Boosting Machines accelerate the learning process and allow for reduced memory usage, particularly due to their use of histogram-based learning.

eXplainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) refers to a set of processes and methods aimed at providing clear and understandable explanations for the decisions offered by Machine Learning models. The architecture of XAI consists of three main components (Adadi and Berrada 2018; Vilone and Longo 2020): 1) Machine Learning Algorithm, 2) Explanation Algorithm 3) Interface The explanation algorithm is used to provide information about the most relevant and influential factors in the process. The interface component presents the information generated by the explanation algorithm.In this study, the two most popular algorithms of Explainable Artificial Intelligence were examined.

Local Interpretable Model-Agnostic Explanations (LIME): LIME is a popular Explainable Artificial Intelligence approach that uses the local behavior of a model to provide interpretable and explainable information about the most relevant and influential factors in predictions. The LIME algorithm generally follows the steps of categorizing numerical variables, generating new observations similar to the dataset's distribution, and developing an explainable model based on this dataset to determine the effects of variables on observations (Ribeiro *et al.* 2016; ElShawi *et al.* 2021). The general mathematical representation of the LIME model is as follows:

$$e(x) = \underset{g \in \varsigma}{\operatorname{argmin}}(f, g, \pi_x) + \Omega(g) \tag{13}$$

In this equation x is the example being explained. The explanation of x (f, g, π_x) is the result of maximizing the fidelity term while considering the complexity $\Omega(g)$.Here, f represents a black-box model being explained, and g represents the interpretable model that explains (Molnar 2018).

Shapley Additive Explanations (SHAP): SHAP is an Explainable Artificial Intelligence approach that uses Shapley values, derived from game theory, to provide interpretable information about the most important and influential factors in predictions. Shapley values originate from cooperative game theory and represent a concept that fairly measures a player's contribution. SHAP provides a framework for understanding how a model makes predictions using these values (Lundberg and Lee 2017; Antwarg *et al.* 2021).The general mathematical formula of the SHAP algorithm is as follows:

$$\phi_i(f) = \frac{1}{N!} \sum_{\pi} \left[f(x_{\pi(i)}) - f(x_\pi) \right] \tag{14}$$

In this equation, $f(x)$ represents the output of the model (where x is the input features.). Here π represents all N! permutations, and $x_{(\pi(i))}$ is thei-th permutation of x's feature according to π . The SHAP value adapts Shapley values to understand the contribution of each parameter to the model's prediction.

Model Performance Metrics

Model performance metrics are measurements used to evaluate the performance of a machine learning model (Cıhan and Coşkun 2021). These metrics can be used to understand how well a model performs, compare different models, or tune the model's hyperparameters.

Confusion Matrix: A confusion matrix is used to interpret the results of a classification model and to cross-examine the errors in the relationship between actual and predicted values.

Table 2 Confusion Matrix showing prediction outcomes versus actual values

		Actual Values	
		Positive (1)	**Negative (0)**
Predicted Values	**Positive (1)**	TP [1,1]	FP [1,0]
	Negative (0)	FN [0,1]	TN [0,0]

- True Positive (TP): Correctly predicting the positive condition.
- True Negative (TN): Correctly predicting the negative condition.
- False Positive (FP): Incorrectly predicting the positive condition.
- False Negative (FN): Incorrectly predicting the negative condition.

Accuracy, Precision, Recall, F1-Score: These scores are derived from the confusion matrix and help provide a clearer understanding of the model's success.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{16}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

$$F1\text{-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

DISCUSSION AND RESULTS

In this study, the operations were performed using Python programming language version 3.11.4. In the dataset used, the predicted parameter was classified based on whether the annual income of a person/observation is less than or greater than \$ 50,000. Among the 34,014 observations, the annual income is less than \$ 50,000, while for 11,208 observations, it is more than \$ 50,000. The percentage distribution of the target variable in the dataset is presented in Figure 1.

Figure 1 Class Distribution of the Target Variable in the Dataset.

After the other parameters in the dataset were adjusted to be suitable inputs for machine learning algorithms, the data was trained using six different machine learning algorithms. Since the target variable in the dataset is binary, classification algorithms were preferred over regression algorithms. The performance metrics of the models used are presented in Figure 2.

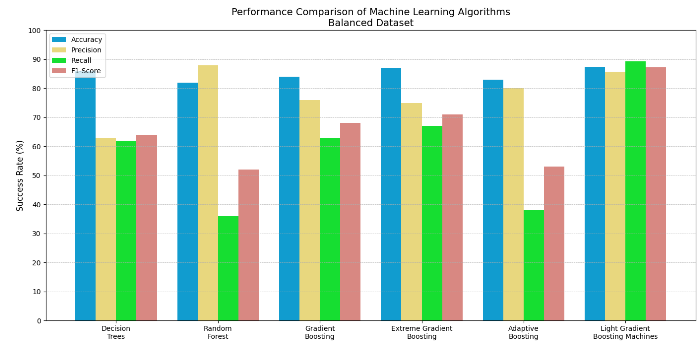


Figure 2 Performance Comparison of Machine Learning Algorithms Balanced Dataset.

Since the target variable's class distribution in the dataset is imbalanced, it is classified as an imbalanced dataset. Performance metrics based on the correct classification rate are unsuitable for this scenario (Chawla *et al.* 2002), as confirmed by the results in Figure 2. To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was applied, generating synthetic examples for the minority class by creating artificial instances along the line segments joining each minority class instance with its nearest neighbors (Blagus and Lusa 2013). After applying SMOTE, the target variable "Income" was balanced, consisting of 23,756 examples earning more than \$ 50,000 and 23,756 examples earning less.

After addressing the imbalance in the dataset, the machine learning algorithms were retrained and tested. The results are presented in Figure 3.

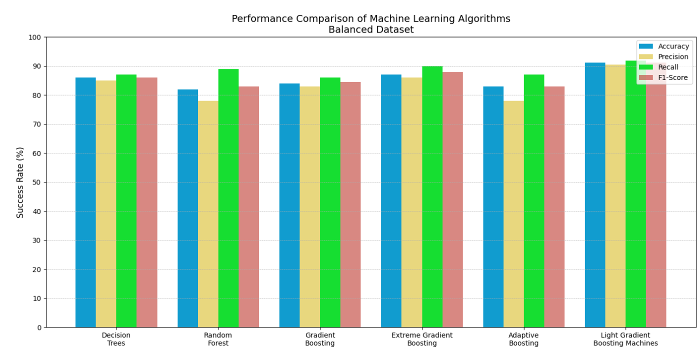


Figure 3 Comparison of Machine Learning Algorithm Performances.

The performance comparisons of the machine learning algorithms used in the study for the imbalanced dataset and the dataset adjusted with SMOTE are shown in Figures 2 and 3. The algorithm that demonstrated the best performance, as seen in these figures, was the Light Gradient Boosting Machines (LightGBM) algorithm. The results obtained from this algorithm were explained using Explainable Artificial Intelligence algorithms.

As shown in Figure 4, the "Capital Gain" parameter has the most significant impact on determining which class a sample belongs to. Following this, "Marital Status" is observed to be another key parameter influencing an individual's annual income. In contrast, the parameters "Workclass" and "Sex" are seen to have the least effect on an individual's annual income.

In Figure 5, we can examine in greater detail the impact of the parameters/variables in the dataset on the classes of the target variable. In the graph in Figure 5a, we observe the SHAP values of the parameters/variables when the annual income of our example

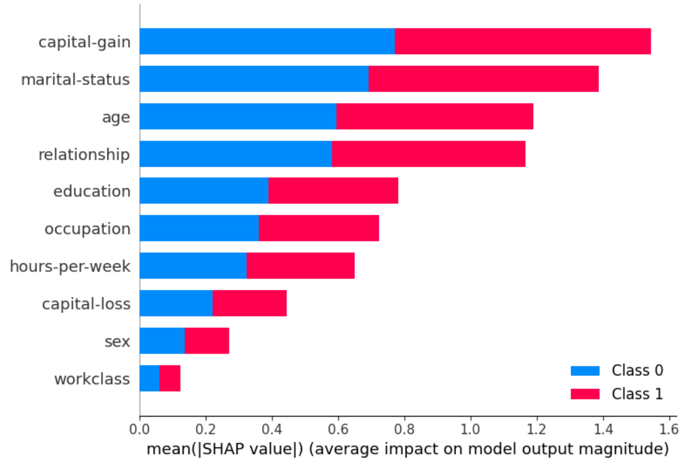
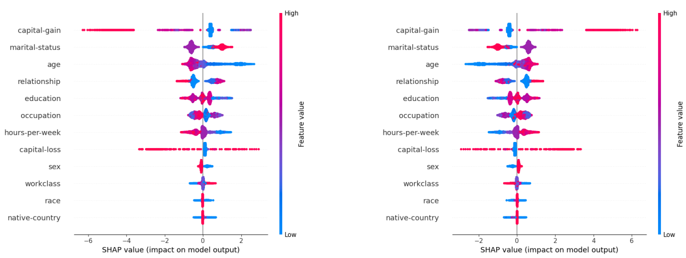


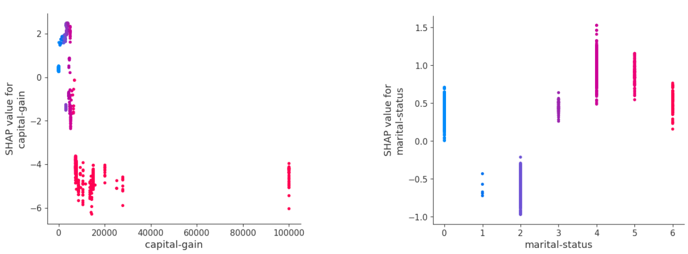
Figure 4 Class Distribution of the Target Variable in the Dataset.



(a) **(b)**

Figure 5 SHAP value distribution for different models. (a) Model A, (b) Model B.

(individual) is less than \$ 50,000, while in Figure 5b, we observe the SHAP values when the annual income exceeds \$ 50,000. For instance, the SHAP value of the "Capital Gain" parameter ranges between -2 and 6 (or between -6 and 2 for the "0" label), and when the target variable's label is 1, the SHAP value ranges from 0 to 6.



(a) **(b)**

Figure 6 Parameter Dependence Plot of the SHAP Algorithm.

Another interface provided by the SHAP algorithm, the Parameter Dependence Plot (Figure 6), allows us to observe the classification of the observation/individual based on the value of the relevant parameter. In Figure 6a, the dependence of the "Capital Gain" parameter on the class to which the observation/individual belongs is shown, while in Figure 6b, the dependence of the "Marital Status" parameter on the class of the observation/individual is illustrated. From this, it can be inferred that if the "Capital Gain" parameter value exceeds 100,000, the annual income of the observation/individual will be more than \$ 50,000. With the SHAP algorithm, an Explainable Artificial Intelligence algorithm, we can clearly observe the impact of parameters/variables and their values on the classes of the target variable. On the other hand, with another Explainable Artificial Intelligence algorithm, LIME, we examine locally which parameters influence the classification of the observation/individual.

(a) **(b)**

Figure 7 Variable Importance Plot of the LIME Algorithm.

The graphs in Figure 7 show the probability of an individual belonging to a specific class, the parameters influencing this classification, and the actual values of those parameters for the observation. For example, the observation/individual in Figure 7a has a 97% probability of having an annual income greater than \$ 50,000. In contrast, the observation/individual in Figure 7b has a 93 % probability of having an annual income less than \$ 50,000.

CONCLUSION

Studies investigating the practical use of machine learning and artificial intelligence algorithms to predict individuals' annual income generally aim to develop income prediction models and achieve more accurate predictions. In line with this goal, this study conducted a comprehensive analysis of several machine learning algorithms using the Python programming language. The imbalanced distribution of the dataset was corrected using the SMOTE method, followed by a comparison of the performance of the machine learning algorithms. At this stage, the best performance was achieved with the Light Gradient Boosting Machines algorithm. Additionally, the impact of parameters/variables on the classes was analyzed using explainable artificial intelligence algorithms. These analyses helped us better understand the study's results and explain the decisions made by the model. These findings underscore the importance of employing effective methods to handle imbalanced datasets and interpret the predictions of machine learning models in data science applications.

Acknowledgments

This work was supported by AI and Data Science Research and Application Center, Sakarya University of Applied Science University Scientific Research Projects Unit.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

Availability of data and material

Not applicable.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

LITERATURE CITED

Adadi, A. and M. Berrada, 2018 Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access **6**: 52138–52160.

Antwarg, L., R. M. Miller, B. Shapira, and L. Rokach, 2021 Explaining Anomalies Detected by Autoencoders Using Shapley Additive Explanations. Expert Systems with Applications **186**: 115736.

Atasoy, N. A. and A. Demiröz, 2021 Makine Öğrenmesi Algoritmaları Kullanılarak Prostat Kanseri Tümör Oluşumunun İncelenmesi. Avrupa Bilim ve Teknoloji Dergisi pp. 87–92.

Becker, B. and R. Kohavi, 2023 Adult, UCI Machine Learning Repository (1996). https://doi.org/10.24432/C5XW20.

Blagus, R. and L. Lusa, 2013 Improved Shrunken Centroid Classifiers for High-Dimensional Class-Imbalanced Data. BMC Bioinformatics **14**: 1–13.

Bulut, F., 2016 Determining Heart Attack Risk Ratio Through AdaBoost. Celal Bayar University Journal of Science **12**: 459–472.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, 2002 SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research **16**: 321–357.

Chen, X., Z. X. Wang, and X. M. Pan, 2019 HIV-1 Tropism Prediction by the XGBoost and HMM Methods. Scientific Reports **9**: 9997.

Cıhan, P. and H. Coşkun, 2021 Performance Comparison of Machine Learning Models for Diabetes Prediction. In *2021 29th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE.

Das, M. and A. Van Soest, 1999 A Panel Data Model for Subjective Information on Household Income Growth. Journal of Economic Behavior & Organization **40**: 409–426.

Dominitz, J., 1998 Earnings Expectations, Revisions, and Realizations. Review of Economics and Statistics **80**: 374–388.

ElShawi, R., Y. Sherif, M. Al-Mallah, and S. Sakr, 2021 Interpretability in Healthcare: A Comparative Study of Local Machine Learning Interpretability Techniques. Computational Intelligence **37**: 1633–1650.

Erdem, F., M. A. Derinpinar, R. Nasirzadehdizaji, O. Y. Selen, D. Z. Şeker, *et al.*, 2018 Rastgele Orman Yöntemi Kullanılarak Kıyı Çizgisi Çıkarımı: İstanbul Örneği. Geomatik **3**: 100–107.

Friedman, J. H., 2001 Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics pp. 1189–1232.

Lazar, A., 2004 Income Prediction via Support Vector Machine. In *Proceedings of the International Conference on Machine Learning and Applications (ICMLA)*, pp. 143–149.

Lundberg, S. and S.-I. Lee, 2017 A Unified Approach to Interpreting Model Predictions. CoRR **abs/1705.07874**: 1–15.

Matkowski, M., 2021 Prediction of Individual Income: A Machine Learning Approach. Preprint or working paper.

Matz, S. C., J. I. Menges, D. J. Stillwell, and H. A. Schwartz, 2019 Predicting Individual-Level Income from Facebook Profiles. PLOS ONE **14**: e0214369.

Michalski, R. S., J. G. Carbonell, and T. M. Mitchell, 2013 *Machine Learning: An Artificial Intelligence Approach*. Springer Science & Business Media.

Mitchell, R. and E. Frank, 2017 Accelerating the XGBoost Algorithm Using GPU Computing. PeerJ Computer Science **3**: 1–20.

Molnar, C., 2018 *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub, https://christophm.github.io/interpretable-ml-book/.

Ribeiro, M. T., S. Singh, and C. Guestrin, 2016 "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, ACM.

Vilone, G. and L. Longo, 2020 Explainable Artificial Intelligence: A Systematic Review. arXiv preprint **arXiv:2006.00093**: 1–39.

How to cite this article: Özkurt, Ö., Garri, F., Yahyaoğlu, B. E., Ağca, O., Bildiren, N. F., and Kaynak, S. Income Level Estimation with Light-GBM:Understanding Model Decisions with Explainable AI Techniques Shap and Lime. *Artificial Intelligence in Applied Sciences*, 1(1), 7-12, 2025.