Income Classification and Regression Models Using Illinois ACS Housing Data

Anthony DiBenedetto
Engineering, Computing, and Mathematical Sciences
Lewis University
Romeoville, IL, United States
anthonypdibenedett@lewisu.edu

Nathaniel D Padal
Engineering, Computing, and Mathematical Sciences
Lewis University
Romeoville, IL, United States
nathanieldpadal@lewisu.edu

Claire Griffin
Engineering, Computing, and Mathematical Sciences
Lewis University
Romeoville, IL, United States
clairekgriffin@lewisu.edu

Brad Baldys
Engineering, Computing, and Mathematical Sciences
Lewis University
Romeoville, IL, United States
bradhbaldys@lewisu.edu

Abstract – This study applies machine learning techniques to predict household incomes using Illinois-specific data from the 2013 American Community Survey. Two models were created: a convolutional neural network (CNN) for regression tasks and a closely mirrored CNN architecture adapted for logistic regression. Data preprocessing involved feature selection, removal of missing values, and exclusion of extreme incomes above $200,000 for the regression model. A $50,000 threshold classified households into low or high income groups for logistic regression. The CNN model, trained with the Adam optimizer and mean squared error loss, achieved a final mean absolute error of 27,832, improving baseline performance by 25%, with an R-squared value of 0.373. The logistic regression model produced an F1-score of 0.77 and an area under the curve of 0.85, demonstrating robust classification performance. These results showcase the potential of machine learning in addressing data gaps and supporting socio economic analysis. Future work could refine the models by incorporating additional features and alternative architectures.

Keywords – Classification, convolutional neural network, income prediction, logistic regression, machine learning.

I. Introduction

The goal of this project is to test the accuracy and capability of neural network models to predict household incomes. Household income demographics are often used by government and company projects to predict public necessities and potential revenue returns. The US and State governments often collect information of homes in the States, which includes the income of the houses themselves. Occasionally, some of the houses are unable to be obtained so being able to get some measure of the predicted income in these homes will allow for more accurate projections. We will be implementing and testing two separate models, a linear regressions and a logistic classification, to achieve this goal of filling in these gaps in the data.

In this report we will look at previous studies into this topic area and then talk about methodology for creating the models. This will include topics about collection and cleaning of data, alongside preliminary analysis and the model creation itself. We can then follow this up with our results from our models and begin to break down our analysis of the models themselves, including effectiveness and reliability. We can question the potential uses of the model and how we could potentially look to improve it with more time. Finally, we will conclude with our overall conclusions and how we feel the results affect our goals for the model, including potential areas for future research and final takeaways.

II. Literature review

The application of machine learning to income prediction has been explored in prior studies, showcasing a variety of methodologies and datasets. Our research builds on this foundation while incorporating unique aspects to advance the field.

Laspiñas and Murcia's study compares six machine learning classifiers, including logistic regression, random forest, and naive bayes, to predict adult income levels [1]. They emphasize the importance of model selection and fine-tuning, with random forest and random tree demonstrating superior performance. Similarly, Jo's research highlights the predictive power of random forest models, particularly after hyperparameter tuning, while evaluating various algorithms, with a focus on predicting whether individuals earn more than $50,000 annually using the UCI Adult Income dataset [2]. Meanwhile, Matkowski explores income prediction using data from the Current Population Survey, focusing on large datasets and machine learning's advantage over traditional methods for improving prediction accuracy in economics [3].

While these studies provide valuable insights into the efficiency of machine learning for income prediction, our research distinguishes itself by focusing on different machine learning model types and less-generalised datasets. Our work integrates unique features such as property value, household compositions, and internet access. This allows for a localized and nuanced understanding of income prediction.

III. Methodology

A. Data

We utilized housing data from the 2013 United States Census [4], focusing on a subset of information specific to Illinois households. After loading the data, we began the cleaning process by selecting columns relevant to our target variable. Guided by personal insight and previous studies, we identified several key attributes for analysis: the number of bedrooms, household size, number of children, property value, land size, house tenure, number of vehicles, household type, household language, and internet access. Some variables were included due to their interconnected nature; for instance, combining the number of bedrooms, household size, and number of children provides a more nuanced understanding of household circumstances. For example, a low bedroom count relative to household size could indicate lower-income situations, as larger households may lack the resources to afford bigger homes. Including attributes like the number of children, often associated with lower income due to higher expenses, helped refine our predictions. This approach aimed to reduce the variability inherent in attributes like household size, ensuring more accurate modeling of income patterns.

After selecting the relevant attributes, the next step was to address missing values and unknown variables in the dataset. An initial review of the cleaned data revealed that approximately 20,000 rows, or 35% of the dataset, had been removed due to missing values. To mitigate this loss, we focused on imputing missing values for key attributes, primarily the property value field. Using a proximity-based imputation approach, we filled missing values by averaging the data from the four closest rows (two above and two below), sorted by the target variable. This method was applied to two attributes, allowing us to recover 10,000 rows of data and effectively halving the data loss. Post-imputation analysis confirmed that the estimated values did not negatively impact model accuracy, ensuring the integrity of the dataset. With this refined dataset, we established a solid foundation for building and training our models.

B. *Model Architecture*

For regression tasks, various modeling options were considered, but we opted for a convolutional neural network (CNN) due to its robustness and ability to handle complex datasets effectively. Simple data preprocessing steps were implemented prior to training the models. For the regression model, households with incomes exceeding $200,000 were excluded to minimize the impact of extreme values. This adjustment was not required for the classification model.

For the logistic classification model, we defined $50,000 as the threshold for categorizing low-income households. This threshold was derived based on the Illinois low-income standards[5], adjusted for the change in income of people in Illinois in 2013 from 2024 [6], and aligned with the average household size in Illinois [7]. These calculations addressed the absence of precise income breakpoints from 2013, ensuring a contextually relevant threshold for classification.

The dataset was split into an 80-20 ratio, dividing it into training and testing sets. During this process, sample weights were derived from the dataset to address representation imbalances. Each household in the dataset included a pre-calculated weight column designed to correct for the over or under-representation of specific groups in the survey. By applying these weights to our machine learning model, we ensured that households with higher weights had proportionally greater influence on the model's predictions.

The CNN architecture begins with an input layer designed to match the size of the dataset, where all features are reshaped into a single channel to ensure compatibility with the CNN. The network comprises four convolutional blocks, each containing two 1D convolutional layers with a kernel size of three and ReLU activation functions to capture spatial relationships between features. Max-pooling with a pool size of two follows each convolutional layer to reduce dimensionality, and batch normalization is applied. After the convolutional blocks, the output is flattened into a 1D vector, which passes through fully connected layers consisting of 512, 256, 128, and 64 neurons, each using a ReLU activation function to introduce non-linearity. Dropout regularization, with a 50% dropout rate, is applied to the larger layers to prevent overfitting. The final layer of the model is either a single neuron dense layer for regression tasks or a dense layer with two or more neurons for classification tasks, depending on the specific application..

For training we used an Adam optimizer and mean squared error (MSE) for the loss function. MSE was used to penalize large deviations from the actual income value, while we used mean absolute error (MAE) as an additional performance metric used to evaluate our models performance. To prevent overfitting, early stopping was implemented and if no improvement was detected for 10 consecutive epochs, training was halted. A batch size of 64 was selected to balance computational efficiency and training stability.

IV. Results

A. *Training*

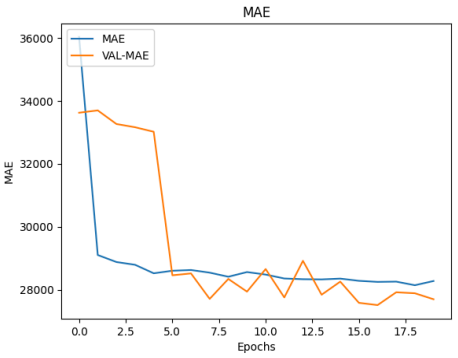

Fig. 1 Epoch MAE loss graph for linear regression.

Figure 1 illustrates a promising trend in our CNN model, with a consistent decrease in MAE across epochs. The simultaneous reduction in training and validation loss suggests that the model is generalizing well to unseen data, indicating minimal overfitting. This training process resulted in a final MAE of 27,832 on the testing set, demonstrating the model's ability to make reasonably accurate predictions.

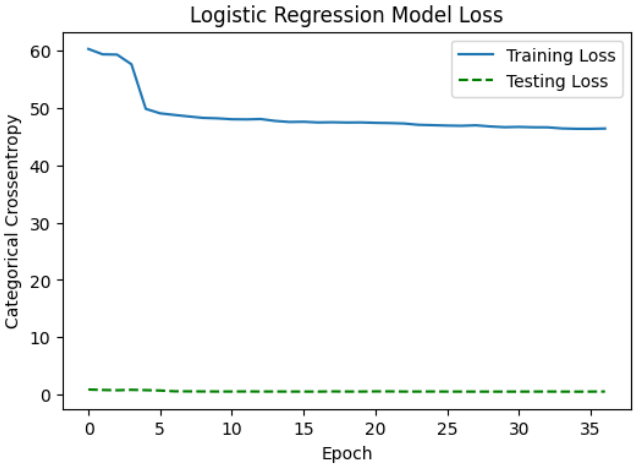

Fig. 2 Epoch loss graph for logistic regression.

B. *Model results*

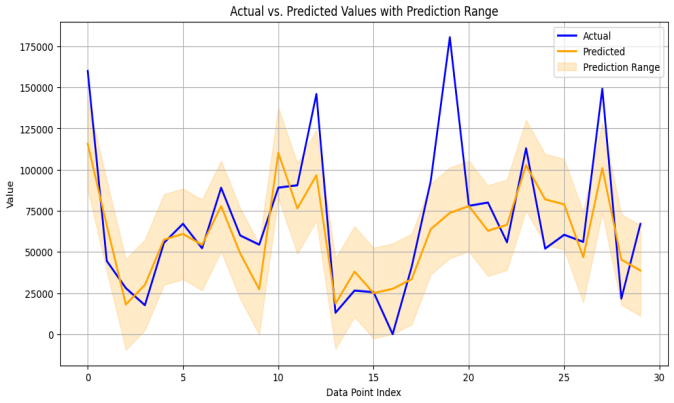

Fig 3. Shown is the performance of the CNN model. Displayed is the comparison of 30 actual values versus the predicted values of household income. There is some noticeable correspondence between the predicted and actual values this graph does show there is room for improvement.

Figure 3 displays the predicted range for each value in the plot, where the final MAE value is added to the predicted values to show their high and low bounds. Visually, the results suggest moderate accuracy, with room for improvement. To better understand the model's performance, we supplemented this visual assessment with additional evaluation metrics.

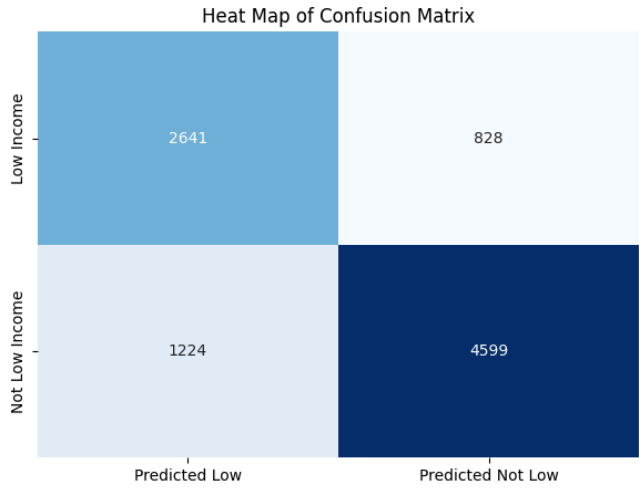

Fig 4. Heatmap of the accuracy of the model, giving the numeric test values for each prediction versus actuality, labeling the density of each section with the color correlated to it.

Our classification model generates a confusion matrix, visualized as a heatmap in Figure 4. This heatmap highlights both the model's successes and its failures. While the model performs well overall, it shows a noticeable bias in misclassifying class 1 more frequently than class 0. Specifically, the test data consists of 40% class 0 and 60% class 1, yet the model exhibits a ~20% higher misclassification rate for class 1. This indicates that the model itself, rather than the dataset, is responsible for this imbalance, as the test data distribution closely mirrors that of the full dataset.

V. Further Analysis

Metric	Value
Mean Target	71470
Mean Prediction	71050
Median Target	63000
Median Prediction	70445
R-Squared	0.373038
MSE	1.28885e+09
MAE	27832
Baseline MAE	37041.5

Table 1. Numerical analysis results performed using the test set.

Table 2 summarizes the key metrics used to evaluate the model's performance. The mean target income of 71,470 and mean prediction of 71,050 indicate that the model's average predictions align closely with the actual average household income, demonstrating reasonable overall accuracy. However, the median target of 63,000 compared to the median prediction of 70,445 suggests a slight overprediction for households near the median income, potentially highlighting challenges in accurately modeling patterns among lower-income households. The R-squared value of 0.373 indicates that the model examples approximately 37.3% of the variance in household income, demonstrating moderate predictive power but also signaling room for improvement. Importantly, the model's MAE of 27, 832 represents a nearly 25% reduction from the baseline MAE of 37,041.5, which assumes the mean target value for all predictions. This improvement underscores the model's ability to effectively learn from the data, despite its limitations.

	Precision	Recall	f1-score	support
0	0.72	0.75	0.73	3709
1	.83	.80	0.81	5583
accuracy			0.78	9292
Macro avg	0.77	0.77	0.77	9292
Weighted avg	0.78	0.78	0.78	9292

Table 2. Confusion matrix results from logistic regression model including f1-scores for more detailed analysis results.

$$F1 = 2 * (Precision * Recall)/(Precision + Recall)$$

Table 2 provides additional evidence that our model exhibits a bias toward one class over the other. The F1 score [6] of 0.73/0.81 shows that while our model

performs reasonably well for real-world applications, it again shows the difference in predicting the two classes. The model has some imperfections, but it is important to note that F1 scores between 0.7 and 0.9 are generally considered acceptable for practical applications, making the model acceptable for many real-world scenarios [8].

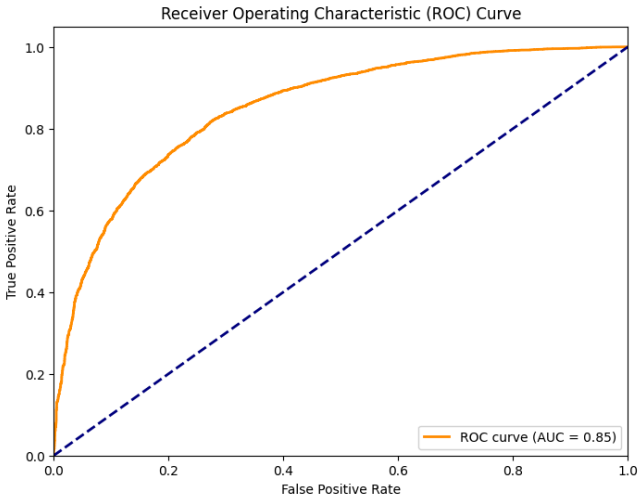

Fig. 5 Receiver Operating Characteristic curve showing AUC for model accuracy for the binary classification.

To further evaluate the model's validity, particularly given its relatively low accuracy for a binary classification model, we created the ROC curve presented in Figure 4. The curve yielded an AUC of 0.85, which is considered a strong score for an ROC curve [9]. This result supports the model's applicability for real-world scenarios and practical implementations, while also leaving room for improvement.

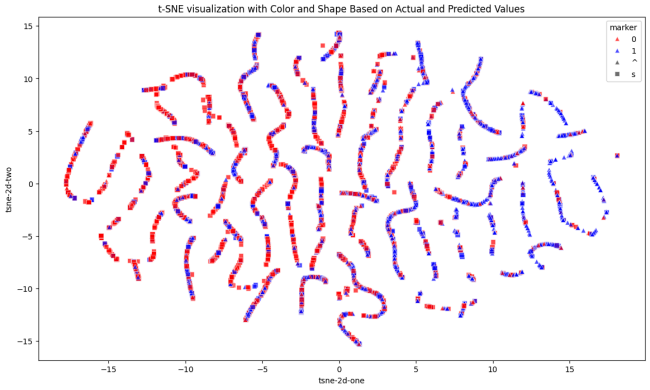

Fig 6. the t-SNE graph to again check the data to ensure that there is some correlation between the classes we are attempting to predict.

As a final check to our model, we created a t-SNE graph or a t-distributed Stochastic Neighbor Embedding graph [10]. This takes a series of data points with similarities and creates string-like datalines on the graph. We can then alter the color and shape of the points to show effectiveness. In Figure 5 we can see that there is a clear divide in the data, in terms of color. One side of the plot is mainly red, while the other is blue, showing that there is some correlation in the data for the model to predict with. We also observed that the dataset contains outliers, with red and blue data points appearing deep within opposing sections, which likely contribute to the error we are experiencing.

VI. Conclusion

A. Future Studies

If this study were to be continued, several potential improvements come to mind. Our analysis utilized only a sample of the full dataset; accessing a more comprehensive dataset with additional features such as geographic location or education levels could help refine the models' results even further. Additionally, given more time, a deeper exploration of feature engineering and analysis could further optimize the performance of the current CNN model. Finally, while this study focused exclusively on CNNs, exploring alternative machine learning models could provide valuable comparisons and potentially yield improved results.

B. Closure

To conclude with our study, we made great progress in our initial goal to use machine learning to predict household incomes from the Illinois ACS housing dataset. By applying both linear regression and logistic classification models, we explored multiple approaches to fill data gaps and provide insights into household income demographics. The results of this study show that while our CNN-based regression and classification models offered reasonable accuracy, there is room for refinement. For instance, the observed bias in class predictions suggests areas where further adjustments, such as rebalancing or feature engineering, could enhance the models' effectiveness. Despite these limitations, the findings reflect the real-world applicability of machine learning tools in fields like public policy, economic forecasting, and social research.

Acknowledgment

We would like to express our gratitude to the Data Science Department for their support and for equipping us with the essential skills and knowledge that were instrumental in conducting this research successfully. We are especially grateful to Dr. Bo Xu for his mentorship and guidance. Finally, we express our thanks to Lewis University for providing us with the opportunity to do this work.

References

[1] E. L. Laspiñas and J. Vianne, "Machine Learning Approaches in Classifying Income Levels," *TWIST*, vol. 19, no. 2, pp. 92–97, 2024, Accessed: Dec. 06,

2024. [Online]. Available: https://twistjournal.net/twist/article/view/214

[2] K. Jo, “Income Prediction Using Machine Learning Techniques,” *escholarship.org*, 2024. https://escholarship.org/uc/item/6d01c9v7

[3] M. Matkowski, “Prediction of Individual Income: A Machine Learning Approach,” *Honors Projects in Economics*, Apr. 2021, Available: https://digitalcommons.bryant.edu/honors_economics/39/

[4] U. C. Bureau, “2013 PUMS Data,” *Census.gov*, Apr. 18, 2024. https://www.census.gov/programs-surveys/acs/microdata/access/2013.html (accessed Dec. 06, 2024).

[5] “IDHS: 01.02.01 - Income Guidelines, 2020-07-01,” *State.il.us*, 2020. https://www.dhs.state.il.us/page.aspx?item=118832

[6] “$1 in 2013 → 2021 | Inflation Calculator,” *www.in2013dollars.com*. https://www.in2013dollars.com/us/inflation/2013?amount=1

[7] “Average size of U.S. households, by state 2018,” Statista. https://www.statista.com/statistics/242265/average-size-of-us-households-by-state/

[8] I. Logunova, “F1 Score in Machine Learning,” Serokell Software Development Company, Jul. 11, 2023. https://serokell.io/blog/a-guide-to-f1-score

[9] J. N. Mandrekar, “Receiver Operating Characteristic Curve in Diagnostic Test Assessment,” *Journal of Thoracic Oncology*, vol. 5, no. 9, pp. 1315–1316, Sep. 2010, doi: https://doi.org/10.1097/jto.0b013e3181ec173d.

[10] K. Erdem (burnpiro), “t-SNE clearly explained,” *Medium*, Apr. 22, 2020. https://towardsdatascience.com/t-sne-clearly-explained-d84c537f53a