Machine Learning Income Prediction

December 10th, 2021

Connor Maas

Table of Contents

1 | ABSTRACT

The relationship between an individual's statistical attributes and financial standing is well-documented. Consequently, by analyzing observable attributes such as occupational information, degree of education, social connections, and demographics, we can make informed predictions about a typically less accessible attribute, *income*. While predicting income may seem superficial at a glance, the resulting implications extend far beyond just the numbers. Income is, in many cases, the determining factor for an individual's societal opportunities, access to resources, and overall quality of life. This paper leverages machine learning classification methods to accurately predict whether an individual's income exceeds or falls below 50,000 USD based on demographic data from the United States Census Bureau. Two classifiers are improved over the course of this paper, Locally Weighted Learning (LWL) and Java RIPper (JRIP). The refined LWL model demonstrates a significant accuracy increase of 5.6% compared to its performance on default settings. Furthermore, the refined JRIP model achieves a final accuracy of 88.8%, marking a nearly 3% improvement over top-performing models evaluated using the same dataset.

2.1 | INTRODUCTION

An individual's income not only reflects their financial status but also their lifestyle in many cases. Excluding substantial inheritances or windfalls like lotteries, income dictates living arrangements, mode of transportation, daily activities, and even specific preferences such as food choices or shopping destinations. Additionally, "greater income is associated with increased happiness and well-being because of the greater advantages" it provides [1]. Since the knowledge of an individual's income provides profound insights into their lifestyle and preferences, the question becomes: is possessing such information beneficial or detrimental?

The answer is both. Entities such as governments or charities can utilize income predictions to aid struggling communities. As an illustration, the United States Census Bureau claims to "calculate poverty status and ask about age and disability status" to help "ensure older people receive appropriate assistance, such as financial assistance with utilities" [2]. Employing the knowledge of income in this way is commendable. Conversely, major corporations are introducing "new ad targeting features which allow marketers to direct ads to people within specific income brackets" [3]. This approach raises concerns as it has the potential to deepen divisions between socio-economic classes.

The primary objective of this paper is to provide readers with methodologies for accurate income prediction, while emphasizing the ethical implications of the insights obtained. Using machine learning classification techniques, I refine two models to predict whether an individual's income exceeds or falls below 50,000 USD.

2.2 | RELATED WORKS

I am not a pioneer in exploring income prediction. In fact, I identified three separate studies that made use of the same dataset, each aiming to classify individuals based on an income threshold of 50,000 USD.

The first study, achieving an overall accuracy of 86%, primarily emphasized data analysis and manipulation [4]. The author, Singh, provided valuable insights into his specific methods for improving prediction accuracy. First, he dedicated a significant portion of his study to examining the data, as he believed "a crucial step before modeling is the exploration of independent variables" [4]. He also noted that "the data needed to be reshaped in order to aid [this] exploration" [4]. His findings clearly indicate that models trained on unrefined data will not yield accurate predictions. In conclusion, it's essential to exercise patience in anticipating results because data requires manipulation before it can be used for training.

The second study, also reporting an accuracy of 86%, concentrated on algorithmic analysis [5]. It specifically stressed the importance of identifying and integrating multiple classifiers. The author, Ra, began his paper by claiming that he had "tested a variety of different models," including "logistic regression, linear discriminant analysis, K-nearest neighbor, classification tree, random forest, and gradient boosting" [5]. Rather than deeply examining the data, Ra opted for a dynamic algorithmic analysis, tweaking his model in response to the outcomes of each experiment. For instance, he found that "despite issues such as the gender wage gap, generated models showed that gender and race were not very significant" factors overall [5]. This discrepancy might have been overlooked during manual analysis, given that we consider race and gender to be significant parts of our identities. While Singh and Ra utilized opposing approaches in building their ultimate models, both were considerably successful.

In the final comparable study I encountered, the reported accuracy was once again 86% [6]. Lichman, the author, emphasized large-scale data structuring. His methodology is outlined as follows: "First, we will split our dataset into two parts - training set and validation set. Then, we will train a model on samples from the training set and finally evaluate its accuracy based on the validation dataset" [6]. A structured approach of this nature enhances the clarity and effectiveness of experiments. Notably, Lichman also highlighted how his model could target "high-income customers exposed to premium products," such as corporations [6]. This aligns with the potential use cases and ethical concerns I discussed in Section 2.1.

In conclusion, while studies have explored numerous methods for income prediction, ranging from ground-level data analysis to algorithmic scrutiny to large-scale data structuring, none of the models have consistently exceeded an accuracy of 86%. A key objective of this paper is to develop a model that surpasses the 86% accuracy threshold.

2.3 | PROCEDURAL OVERVIEW

- Section 2 encompasses an introduction to the topic, the key objectives of this paper, an acknowledgement of prior research, and a brief overview of the procedures.
- Section 3.1 provides detailed information about the origins of the dataset.
- Section 3.2 outlines the dataset's features, data manipulation, and dataset partitioning.
- Sections 3.3 and 3.4 cover data analysis and subsequent classifier selection, using the machine learning platforms Weka and Lightside.
- Section 4 presents the baseline performances of the chosen classifiers.
- Sections 5.1 and 5.2 are dedicated to error analyses and parameter tuning, both of which play a pivotal role in enhancing the model's accuracy.
- Section 6 describes final testing on the dataset, followed by a discussion of this paper's findings in Section 7.

3.1 | DATA COLLECTION

The dataset is sourced from authentic 1994 census data, made publicly available by the United States Census Bureau. However, it was Barry Becker, the General Manager for the United States Alliances and Partnerships, who compiled the dataset into its final form [8]. According to the Census Bureau, the data "comes from a variety of sources," including but not limited to anonymous surveys, governmental and commercial entities, and online databases [9]. While the accuracy of Census Bureau's data might waver in certain instances, such as population estimations, this dataset primarily consists of straightforward recorded data with minimal forecasting. The University of California Berkeley wrote an article about the reliability of the Census Bureau, claiming it is "one of the most reliable sources of U.S. demographics statistics and data" [10]. This assertion gains weight when considering the limited number of sources that have access to massive amounts of data on demographics and income. In conclusion, this dataset is a reputable resource for analytical purposes.

3.2 | DATA PREPARATION

My first course of action was to randomly[1] select instances from the dataset, reducing it in size for computational efficiency. Initially, there were 32,562 instances, but this number was reduced to 1,250. Table 1 on the next page presents the features, accompanied by types, explanations, and examples.

[1] I assigned a random number to each instance, sorted these instances by their assigned number, and selected the first 1,250.

Table 1:

Feature	Type	Explanation	Examples
age	Numeric	The number of years an individual has lived	43, 82
workclass	Nominal	A general term used to categorize occupations	private, self-employed-not-inc
fnlwgt	Numeric	The number of people which an individual is believed represent according to census data	54878, 7322, 588932
education	Nominal	The highest level of education which an individual has completed	HS-grad, bachelors
education-num	Numeric	A numeric representation of the education feature	6, 9, 13
marital-status	Nominal	A classification for an individual's marriage history	divorced, married-civ-spouse
occupation	Nominal	A specific category describing an individual’s occupation	sales, prof-specialty
relationship	Nominal	An individual's relationship status from the perspective of others	husband, not-in-family, own-child
race	Nominal	A classification for individual's racial background	black, white
sex	Nominal	The biological gender of an individual	male, female
capital-gain	Numeric	USD amount of profit from investments	0, 7688
capital-loss	Numeric	USD amount of loss from investment	0, 1887
hours-per-work	Numeric	The number of hours an individual works per week	45, 60, 65
native-country	Nominal	An individual’s native country (place of birth)	united-states, iran, mexico
income (class value)	Nominal	An individual’s yearly income	<=50k, >50k

As discussed in Section 2.1, income serves as a valuable statistic. Given both its significance and nominal nature in this dataset, it's a suitable choice for a class value. As illustrated in Table 1, eight of thirteen features are nominal. Due to platform specific behavior within both Lightside and Weka, an abundance of nominal features can complicate the process of enhancing a model's accuracy. Therefore, I opted to transform these nominal features into binary features, yielding 97 numeric features as replacements. Subsequently, I standardized the initial numeric features to facilitate an accurate comparison of average cell values during my error analysis, as detailed in Section 5.1. Sample instances of the manipulated data are presented in Table 2 below.

Table 2:

Feature	**Type**	**Explanation**	**Example**
'workclass = private'	Numeric	1 if workclass is private, else 0	0, 1
'workclass = State-gov'	Numeric	1 if workclass is private, else 0	0, 1
'education = masters'	Numeric	1 if workclass is private, else 0	0, 1
age (standardized)	Numeric	Standardized values of age	-0.026619, 0.73729
fnlwgt (standardized)	Numeric	Standardized values of fnlwgt	0.943722, - 0.13139

The next step in this process involved randomly partitioning both the original and transformed datasets into development sets (20% of instances), cross-validation sets (70% of instances), and final testing sets (10% of instances). I used the development set for technical analysis and model testing, while the cross-validation set was employed for model training and parameter tuning. To ensure the trained model remained unbiased, the final testing set was withheld from all training or development processes.

For an effective parameter tuning procedure, detailed in Section 5.2, it's essential to evenly divide the binary cross-validation set into ten distinct subsets. Among these subsets, five serve as training sets, while the other five are designated as testing sets. Each training set is used to train a model, which is then tested on its corresponding testing set. I made use of Weka's unsupervised filter, "removeFolds," to facilitate this division, segregating the data by fold into respective testing and training sets.

The datasets mentioned above are illustrated in Figure 1 on the next page, which I reference accordingly in the proceeding sections.

Figure 1:

| Complete dataset *(Dataset 0)*
    | Reduced dataset *(Dataset 1)*
        | Development set *(Dataset 1B)*
        | Final testing set *(Dataset 1F)*
    | Reduced dataset (transformed) *(Dataset 2)*
        | Development set (transformed) *(Dataset 2D)*
        | Cross validation set (transformed) *(Dataset 2C)*
            | Training 1 - 5 sets (transformed) *(Datasets 2CTr 1 - 5)*
            | Testing 1 - 5 sets (transformed) *(Datasets 2CTe 1 - 5)*
        | Final testing set (transformed) *(Dataset 2D)*

Note: An indented set X under another set Y indicates that X is a subset of Y.

3.3 | CLUSTER ANALYSIS

I performed a cluster analysis on Dataset 1B (Figure 1) to extract preliminary insights about the data, enabling justified classifier selections in Section 3.4. I first experimented with various cluster counts, observing that an increasing number of clusters only reduced the exclusivity of each cluster. Given that eight features were nominal, I then performed a classes-to-clusters[2] analysis for every cluster count. As expected, a rise in the number of clusters led to an increase in mis-clustered instances. The performance is displayed in Table 3 below.

Table 3:

Number of Clusters	Incorrectly Classified Instances
1	26.4
2	37.6
3	50
4	60.4
5	64.8
6	72.8
7	73.6
8	74.8

[2] A method which clusters without class attributes and then assigns classes to clusters based on majority.

As indicated in Table 3, using a single cluster appears to be the most effective method for grouping the data. However, it's important to note that 73.6% of the instances in this dataset fall within the "<=50k" income category. As a result, the model could potentially classify all instances as "<=50k" income and still attain a deceptively high accuracy. Ultimately, I opted for two clusters as it provides insights into the data while still ensuring the majority of instances are clustered correctly.

Table 4 below showcases three random instances from each cluster. Given that there were 13 features, I trimmed the table to include only the relevant ones for comparison with the cluster instances' centroid.

Table 4:

Age	Workclass	Education	Marital Status	Occupation	Sex	Income	Cluster
24	Private	7th-8th	Never	handlers-cleaners	M	<=50k	1
26	Private	Some-college	Never	adm-clerical	M	<=50k	1
35	Private	HS-grad	Never	exec-managerial	F	<=50k	1
43	Private	Bachelors	Married	prof-specialty	M	>50k	2
57	Self-employed	Prof-school	Married	prof-specialty	M	>50k	2
44	Self-employed	HS-grad	Married	exec-managerial	M	<=50k	2

Most features align with the centroid instances of the cluster to which they were assigned. In all instances of Cluster 1, the working class was "private," and the marital status was "never married." In multiple instances of Cluster 1, the education level was "some-college," the occupation was "adm-clerical," and the sex was "male." For all cluster 2 instances, the sex was "male" and the marital status was "married." In some instances of Cluster 2, the working class was "private," the occupation was "prof-specialty," and the education level was "HS-grad."

Both Cluster 1 and 2 are fairly homogeneous. Specifically, I noticed that: (1) Cluster 2 instances were generally comprised of older individuals as compared to Cluster 1, (2) Cluster 2 instances had a higher level of education than Cluster 1 instances, (3) Cluster 1 instances were predominantly unmarried, while Cluster 2 instances were all married, (4) Cluster 2 included two individuals with professional specialties as their occupation and two self-employed individuals, whereas Cluster 1 had neither, and finally, (5) only Cluster 1 had instances that included a female. Based on these instances, I mentally grouped the data into two categories:

- Cluster 1: “Young, less educated, unmarried males and females”
- Cluster 2: “Older, more educated, professional, married males”

3.4 | CLASSIFIER SELECTION

A key objective of this paper, as elaborated on in Section 2, is to develop a model that achieves maximal accuracy. Hence, it's crucial to select classifiers that are not only accurate but also capable of improving.

I started by placing instances into context, viewing them as individuals rather than mere data points. In doing so, it quickly became apparent that income classification would invariably have outliers. This stems from the fact that individuals aren't aggregates of values. Rather, each person possesses the immeasurable ability to influence their own outcomes, regardless of their demographics or background. For instance, in Section 3.3 (Table 4), an executive manager, married, male was noted to have an income of less than 50,000 USD annually, which contradicted the conclusions drawn in that section.

Based on the above observation, I chose the Locally Weighted Learning (LWL) classifier, as it excels at disregarding anomalies and prioritizing the proximity of instances. Nonetheless, I was skeptical about disproportionate weighting of certain categories. Having identified six significant features in Section 3.3, my goal was for the model to grasp the broader context instead of fixating on a few heavily weighted instances. I also aimed to prevent the model from treating every category with equal importance. Therefore, I adopted the WeightedInstancesHandlerWrapper (WIHW[3]) to better manage weighted instances.

I also predicted the Java RIPper (JRIP[4]) classifier would be successful, given that rule-based classification is likely to be effective in interpreting the diverse set of binary categories in Dataset 2C (Figure 1). For example, as suggested in Section 3.3, a particular combination of categories such as “older, more educated, professional, married males” might be formulated as a rule.

Finally, I evaluated the performances of the Naive Bayes and Logistic Regression classifiers to better understand how my selected classifiers compared. The performance of the classifiers on default settings using a ten-fold cross-validation over Dataset 2D is displayed in Table 5 below.

Table 5:

Classifier	Correctly Classified Instances
JRIP	82.9714%
Naive Bayes	82.8571%
Logistic Regression	80.8%
LWL (WIHW)	75.0857%

[3] A tool that enhances a model's ability to consider data with differing levels of importance.
[4] A classifier which uses ordered rule lists and reduced-error pruning.

Table 5 revealed that the JRIP classifier outperformed the remaining three, leading me to select it as a classifier for this paper. Conversely, LWL (WIHW) was significantly less effective as compared to the more established classifiers, Naive Bayes and Logistic Regression. However, based on practical observations of the data and results of the cluster analysis, I still opted to move forward with improving this model. I suspected that the default settings hindered LWL (WIHW) from achieving a competitive performance. In conclusion, I proceeded in an attempt to improve both the JRIP and LWL (WIHW) classifiers.

4 | BASELINE PERFORMANCES

The JRIP and LWL (WIHW) classifiers were trained using Dataset 2C (Figure 1) and tested on Dataset 2D (Figure 1). The results are presented in Table 6 below.

Table 6:

Classifier	Correctly Classified Instances	Kappa Statistic
JRIP	86%	0.5905
LWL (WIHW)	76%	0

I made two observations given the baseline results. (1) The Kappa statistic for LWL (WIHW) is 0, indicating that the model predicts every instance as a single value, either "<=50k" or ">50k." (2) JRIP's performance increased significantly in comparison to Section 3.4.

5.1 | ERROR ANALYSIS

I utilized the Support Vector Machine (SVM) classifier in Lightside to conduct my error analysis and swiftly pinpointed four problematic features. The initial performance of SVM is presented in Table 7 below.

Table 7:

Classifier	Correctly Classified Instances	Kappa Statistic
Support Vector Machines	0.808	0.4676

Error 1:

I identified "education = HS-grad" as the first problematic feature, as it often led the model to incorrectly predict an income of "<=50k" instead of the correct income of ">50k." The

feature exhibited a horizontal absolute difference of 0.2011, a mean cell value of 0.2979, and was assigned a weight of 0.9089.

Logically, one would assume the presence of "education = HS-grad," which indicates the absence of a college degree, would suggest a lower income due to fewer opportunities for higher-paying employment. However, this wasn't necessarily the case. I hypothesized that identifying and excluding a particular demographic of outliers could refine the model's predictions.

Upon analyzing the data, I observed that 10 out of 12 individuals with an income of ">50k" were married. Among those in the other class, that number was 26 out of 57. Additionally, 11 out of 12 individuals with an income of ">50k" were male, while the sex distribution was more balanced for the other class. Finally, all individuals with an income of ">50k" were identified as white in terms of race. Based on these observations, being a married, white male was a strong indicator that an individual had an income of ">50k."

To address the problematic feature, I introduced a new binary feature tailored to the above demographic, termed "specific-HS-grad." Figure 2 below presents an interpretation of this new feature in the form of pseudocode.

Figure 2:

```
IF education = HS-grad
        AND marital-status = married-civ-spouse
        AND race = white
        AND sex = male
THEN: 1
ELSE: 0
```

After incorporating "specific-HS-grad," I re-evaluated the model. The updated performance is presented in Table 8 below.

Table 8:

Classifier	Correctly Classified Instances	Kappa Statistic
Support Vector Machines	0.8286	0.5243

I then compared this performance with the baseline to gauge improvement. Given that $0.05 < p^5 < 0.1$, the model showed marginal improvement. Therefore, this discovery offered a solution to the identified error and held potential for enhancing the performance for our selected classifiers. I intended to incorporate this feature into the dataset for subsequent evaluations.

[5] A numeric value that when < 0.05 indicates "statistical significance."

Error 2:

I identified “hours-per-week” as the next problematic feature, as it caused the model to mispredict incomes of “>50k” when the correct income was “<=50k”. The feature displayed a significant horizontal absolute difference of 1.0001, with an average cell value at 0.7689 and a feature weight of 0.3749.

It's a common assumption that individuals working longer hours earn more money. However, the type of occupation frequently plays a more significant role in determining an individual's earnings. The model struggled to accurately predict the incomes of a subset of individuals who worked fewer hours but earned “>50k.” My goal was to accurately classify this particular group by using their professions as stronger indicators.

After some manual data analysis, I observed that 15 out of 25 individuals who worked 40 hours per week and earned “>50k” had occupations listed as either “prof-specialty” or “exec-managerial.” Conversely, those earning “<=50k” were much less likely to hold these roles.

To address this error, I added an additional 10 hours to the “hours-per-week” column for individuals in professional specialties or executive manager roles. The updated category is presented as pseudocode in Figure 3 below.

Figure 3:

```
IF occupation = prof-specialty
        OR occupation = exec-managerial
THEN: hours-per-week + 10
ELSE: hours-per-week
```

I subsequently standardized the new feature for simplified comparison with the other standardized features in the dataset. The model's performance, after incorporating the new feature, is detailed in Table 9 below.

Table 9:

Classifier	Correctly Classified Instances	Kappa Statistic
Support Vector Machines	0.7943	0.3965

The new model did not improve as I had anticipated and, in fact, decreased in both accuracy and Kappa value. Statistically, however, the difference between this model and the baseline model was insignificant, producing a p-value of 0.28. In summary, both the error analysis and proposed solution for this feature proved to be ineffective. I chose not to incorporate the new feature in future iterations, as it would be unlikely to enhance the model's overall accuracy.

Error 3:

I identified “marital status = married-civ-spouse” as a problematic feature, as it caused the model to frequently mispredict an income of “>50k” when the correct income was “<=50k.” The feature exhibited a vertical absolute difference of 0.0018, an average cell value of 0.8727, and a feature weight of 0.8254.

Interpreting this error was challenging because I had previously concluded that in most instances, being married was a strong indicator of a “>50k” income. However, the low horizontal absolute difference suggested otherwise, forcing me to look deeper into the data to understand this anomaly.

I noticed that married individuals under 60 years old, with either capital gains or capital losses, tended to have incomes exceeding “>50k.” Perhaps, such features might correlate with higher incomes because a greater risk tolerance, indicated by capital gains and losses, can sometimes yield larger financial successes. However, the exact reasons for this correlation remained beyond my immediate understanding.

To address the error, I introduced a new binary feature called “exceptional-marriage,” which I set to “1” for individuals under 60 years old with either capital gains or losses. This category is illustrated as pseudocode in Figure 4 below.

Figure 4:

```
IF marital-status = married-civ-spouse
        AND age < 60
        AND (capital-gains > 0 OR capital-losses > 0)
THEN: 1
ELSE: 0
```

Upon incorporating this new feature, I re-evaluated the model. The updated performance is displayed in Table 10 below.

Table 10:

Classifier	Correctly Classified Instances	Kappa Statistic
Support Vector Machines	0.8126	0.4646

I then contrasted this model against the baseline to assess the impact of these additions on performance. Given that $p = 0.696 > 0.05$, the model’s improvement was not statistically significant. However, I noted that both the accuracy and Kappa value increased slightly. I decided to proceed with caution when incorporating this feature, anticipating that its minor enhancement might prove to be beneficial. Additionally, I theorized that its effectiveness might be more pronounced in other subsets of the data.

Error 4:

The final problematic feature I identified was “relationship = husband,” as it caused the model to frequently mispredict incomes as “>50k” when the correct prediction was “<=50k.” The feature exhibited a vertical absolute difference of 0.0575, an average cell value of 0.8025, and a feature weight of 0.1398.

I first noticed that the “marital-status” and “relationship” features were inherently connected. Specifically, if someone was labeled as a “husband,” it suggested that they must be married, and vice versa. When converting nominal features to binary, each value of the nominal feature became its own distinct binary feature. This meant that the value “relationship = husband” was being evaluated separately, even though its implication was already captured in the “marital-status” feature. Retaining “relationship = husband” in the dataset was therefore redundant and put too much weight on this feature.

As a result, I decided to remove “relationship = husband” from the dataset completely. I then re-evaluated the model on the updated dataset. The performance is displayed in Table 11 below.

Table 11:

Classifier	Correctly Classified Instances	Kappa Statistic
Support Vector Machines	0.8137	0.467

I compared the proposed model to the baseline to gauge the performance difference. Given that $p = 0.696 > 0.05$, the improvement in the model was, again, not statistically significant. However, I chose to cautiously incorporate this feature into the dataset based on the rationale outlined in Error 3.

Evaluation of Error Analysis:

Using the results of the SVM error analysis, I further investigated the JRIP and LWL (WIHW) models. I observed that the JRIP model exhibited the same problematic features as SVM, namely “marital status = married-civ-spouse,” “hours-per-week,” “education = HS-grad,” and “relationship = husband.” This strongly indicated that these features were problematic for various classifiers. Since the LWL (WIHW) model classified all instances as a single value, pinpointing specific problematic features became challenging. However, given JRIP's insights on certain features being problematic in general, it's reasonable to infer that these features might have influenced the LWL (WIHW) model's performance as well. In conclusion, I decided to implement the following changes: (1) add the feature “specific-HS-grad,” (2) add the feature “exception-marriage,” and (3) remove the feature “relationship = husband.”

5.2 | PARAMETER TUNING

I fine-tuned both the JRIP and LWL (WIHW) models using the Weka Experimenter. Specifically, I applied varied parameter configurations during training on Datasets 2CTr 1 - 5 (Figure 1), and evaluated their subsequent performances on Datasets 2CTe 1 - 5 (Figure 1).

JRIP Tuning:

Using default settings, the JRIP model achieved a baseline performance of 83.84% correctly classified. Since JRIP had two parameters to tune, namely "minNo" and "optimizations," numerous successful combinations were possible. Upon testing various combinations of these two parameters, I observed a decline in the JRIP model's performance as "minNo" increased. Consequently, I retained "minNo"'s default value of 2.0 and concentrated on fine-tuning the "optimizations" parameter. I undertook a standard tuning procedure, as shown in Table 13, by adjusting the optimization settings to identify the top-performing model.

Table 12:

Test = Test set % correct
Opts = Training set % correct (optimizations = x, minNo = 2.0)

Fold	Opts = 2	Opts = 10	Opts = 50	Opts = 100	Opts = 500	Best Setting	Test
1	83.14	83.71	83.71	84.00	84.57	500	85.1429
2	82.14	81.43	81.43	81.43	81.43	2	83.4286
3	81.29	83.00	83.57	83.29	83.14	50	85.1429
4	84.71	86.43	86.29	86.14	86.57	500	80
5	83.43	83.57	84.14	84.14	83.29	50	84.5714

Average = (85.1429 + 83.4286 + 85.1429 + 80 + 84.5714) / 5 = **83.65716 correctly classified**

"Optimizations = 50" and "optimizations = 500" appeared most frequently as the best setting. I selected "optimizations = 50" as the preferred setting due to its repeated occurrences and quicker execution time relative to "optimizations = 500." Although tuning didn't yield a significant statistical improvement, as $p^6 > 0.05$, I integrated the "optimizations = 50" setting into my final model due to the observed minor enhancement.

[6] Ibid.

LWL (WIHW) Tuning:

Using default settings, the baseline performance of the LWL (WIHW) model produced a classification accuracy of 75.36% Given LWL (WIHW) only has one tunable parameter, I bypassed experimentation and immediately proceeded with a standard parameter-tuning procedure. The results are as shown in Table 13.

Table 13:

Test = Test set % correct
KNN = Training set % correct (KNN = x)

Fold	KNN = 0	KNN = 10	KNN = 100	KNN = 500	Best Setting	Test
1	74.59	78.40	80.95	74.59	100	82.8571
2	75.48	78.15	81.57	75.48	100	80
3	74.97	77.64	82.33	74.97	100	82.2857
4	75.48	77.14	81.84	74.46	100	76
5	74.46	78.27	80.93	74.46	100	83.8571

Average = (82.8571 + 80 + 82.2857 + 76 + 83.8571) / 5 = **80.99998% correctly classified**

Therefore, tuning this model yielded significant results, as $p < 0.05$. Given the statistically significant improvement, I integrated the optimal setting of "KNN = 100" into the final model.

6 | FINAL RESULTS

I trained over Dataset 2C (Figure 1) with the JRIP and LWL (WIHW) models on default settings. I then tested on Dataset 2D (Figure 1). The results of this evaluation are presented in Table 14 below.

Table 14:

Classifier	Correctly Classified Instances	Kappa Statistic
JRIP	84%	0.4843
LWL (WIHW)	80.8%	0

I then modified the two datasets excluding the “relationship = husband" feature, and adding the “specific-HS-grad” and “exception-marriage” features discussed in Section 5.1. I then trained the JRIP model over Dataset 2C (Figure 1) with “optimizations = 50” and “minNo = 2.0.” I did the same for the LWL (WIHW) model with “KNN = 100.” Finally, I used Dataset 2D (Figure 1) as the test set to obtain the final results, presented in Table 15 below.

Table 15:

Classifier	Correctly Classified Instances	Kappa Statistic
JRIP	88.8%	0.6501
LWL (WIHW)	86.4%	0.5068

For both the JRIP and LWL (WIHW) models, the performance increased with statistical significance, as $p^7 < 0.05$.

7 | DISCUSSION

In this paper, I outlined the development of two distinct models capable of accurately predicting whether an individual's income exceeds or falls below 50,000 USD. Beginning with a brief overview of the topic’s significance, I then dove into data analysis and processing. Based on this analysis, I chose to assess and enhance two classifiers: Locally Weighted Learning (LWL) utilizing WeightedInstancesHandlerWrapper (WIHW) and Java RIPper (JRIP). Subsequently, I carried out baseline experiments, undertook multiple error analyses, and employed parameter tuning for model optimization. Both models ultimately demonstrated statistically significant improvements. Notably, in Section 2.2, I aimed to surpass 86% classification accuracy, a benchmark that three referenced papers using the same dataset could not attain. The final JRIP model recorded an accuracy of 88.8%, marking an improvement of almost 3% when compared to this benchmark. The refined LWL (WIHW) model, which achieved an accuracy of 86.4%, saw substantial improvement from its performance of 80.8% using default settings.

In Section 2.2, I discuss the methodologies that were previously utilized for modeling this dataset. Both the JRIP and LWL (WIHW) models were significantly influenced by these approaches, with two strategies standing out as particularly impactful. Firstly, Lichman's recommendation for strategic data partitioning ensured robust training and testing results. Secondly, Singh's emphasis on detailed data analysis unveiled several unexpected findings. Notably, in Section 3.4, I chose to further refine the LWL (WIHW) model, even though JRIP, Naive Bayes, and Logistic Regression demonstrated higher accuracy. This decision was influenced by my detailed examination of the data in Section 3.3 during cluster analysis. I recommend that future researchers adopt a comparable approach when analyzing similar datasets.

[7] Ibid.

I observed that parameter tuning yielded exceptional improvements, in contrast to the error analysis. The maximum improvement I observed during error analysis was statistically marginal. However, the tuning process led to statistically significant performance improvements across various parameter settings. In particular, the performance of the LWL (WIHW) model saw a significant boost, which I attribute primarily to the tuning. While the JRIP model demonstrated a statistical improvement, it was not nearly as pronounced as that of the LWL (WIHW) model. This is likely a result of JRIPs inherent effectiveness as a classifier, which poses challenges for making substantial enhancements. In conclusion, tuning proved to be the superior method and should be evidently prioritized when analyzing data pertaining to individual profiles, backgrounds, and demographics.

A limitation of this paper is that the discussed models can't be used to effectively classify individual instances of the data. In other words, the models are not designed to determine the income of a singular person, but rather excel at categorizing groups of people into labels such as "largely <=50k" or "predominately >50k." Another consideration is the limited availability of datasets that encompass both an individual's income and demographics. Consequently, sourcing new data that can substantially enhance the model's accuracy is challenging.

I advocate for future research to focus on broadening the practical uses of these classifiers as opposed to further refining model accuracy. As highlighted in Section 2.1, possessing the knowledge of an individual's income carries significant weight. Therefore, it's imperative to exercise its applications with responsibility.

8 | REFERENCES

[1] Matta, Christy. "How Does Your Income Determine Your Well-Being?" *Mental Help How Does Your Income Determine Your WellBeing Comments*, https://www.mentalhelp.net/blogs/how-does-your-income-determine-your-well-being/.

[2] Bureau, US Census. "Why We Ask about... Income." *Why We Ask About... Income | American Community Survey | US Census Bureau*, https://www.census.gov/acs/www/about/why-we-ask-each-question/income/.

[3] "Facebook Introduces Household Income Targeting Based on U.S. ZIP Code Averages." *MarTech*, 25 Oct. 2021, https://martech.org/facebook-introduces-household-income-targeting-based-on-u-s-zip-code-averages/.

[4] Singh, Author Abhinav. "Predicting Income Level, an Analytics Case Study in R." *CloudxLab Blog*, 13 Sept. 2017, https://cloudxlab.com/blog/predicting-income-level-case-study-r/.

[5] youngcr003. "Predicting Income Using US Census Data." *Kaggle*, Kaggle, 15 Sept. 2017, https://www.kaggle.com/youngcr003/predicting-income-using-us-census-data.

[6] "Income Prediction." *Income Prediction*, https://seahorse.deepsense.ai/casestudies/income_predicting.html.

[7] Datopian. “Adult Income.” *DataHub*, https://datahub.io/machine-learning/adult#data.

[8] *Barry Becker - General Manager - US ... - Linkedin.cn*. https://www.linkedin.cn/in/barry-becker-25a2718.

[9] Bureau, US Census. “Combining Data – a General Overview.” *Census.gov*, 18 Nov. 2021, https://www.census.gov/about/what/admin-data.html.

[10] Lee, Corliss. “Who, How Many and Where: Research Using the U.S. Census.” *UC Berkeley Library Update*, 8 Oct. 2016, https://update.lib.berkeley.edu/2016/05/04/who-how-many-and-where-research-using-the-u-s-census/.