# Income Evaluation

**Dylan Mather, Layth Zubaidi**

CS 577
San Diego State University
December 13th 2022

# 1 Abstract

For our research project we ran various machine learning models to predict if an individual's annual income was above or below 50K. The dataset includes an individual's education level, age, gender, occupation, etc which can all play a factor in determining how much money they make. This data was collected by the U.S Census Bureau from 1994. The goal of this project was to determine what factors that we can control in our life would have the greatest influence on the amount of money we make. We have applied machine learning algorithms to this dataset in a Jupyter Notebook to conduct an analysis of the variables to find those that are highly correlated to wealth. We implemented a test/train split on our data so we could measure the accuracy of each of the models we tested. To further decrease the variance created by the random split, we also conducted cross validation for each of the models. Then we implemented featured engineering to our dataset in the form of One Hot Encoding. We have manipulated the data in different ways, such as removing unnecessary/unneeded columns, normalizing of the data points, as well as scaling the data. We collected multiple metrics for each models as well as creating visualizations to explore our results. We ultimately found that our Logistic Regression model, based on our AOC metric, produced the best results. The features that had the greatest influence on the models was the age and the level of schooling completed. The data source comes from online, downloaded from Kaggle.com (https://www.kaggle.com/datasets/wenruliu/adult-income-dataset).

# 2 Introduction

With a multitude of different datasets one can pick and analyze many different subjects concerning any social, economic, or political topic. A dataset that focuses on an individuals success we found to be an interesting topic to analyze about multiple attributes of an individual and find if there are key points that can lead one to making an adequate living. Analyzing census data is critical for developing accurate assessments of economic well-being for the Nation as a whole as well as for different racial, ethnic, and gender populations. Income statistics enable us to know about the distribution of income for a given population. The prominent inequality of wealth and income is a huge concern, especially in the United States. The likelihood of diminishing poverty is one valid reason to reduce the world's surging level of economic inequality. The principle of universal moral equality ensures sustainable development and improves the economic stability of a nation. We are curious about what variables have the biggest influence on an increase or decrease in income. It is important to analyze the data to compare the income of the people who belong to different sectors, countries, and occupations. The performance can be evaluated gender-wise and age wise. It primarily aims at learning the various factors that can help our evaluation process of what determines a salary less or greater than 50,000. The information provided in the income evaluation dataset can be used to give better services, improve the quality of life, and identify the problems and solutions of a population. The fields of Data Analysis lets us explore certain hidden patterns and concepts which can lead to the prediction of future events. The problem of income inequality has been of great concern in recent years. Making the poor better off does not seem to be the sole criteria to be in the quest for eradicating this issue. People of the United States believe that the advent of economic inequality is unacceptable and demands a fair share of wealth in society. This model aims to conduct a comprehensive analysis to highlight the key factors that are necessary for improving an individual's income. Such an analysis helps to set focus on the important areas which can significantly improve the income levels of individuals.

To approach this problem we wanted to use many and different kinds of machine learning algorithms, as well as visual analysis on the models and the columns of data points themselves. Of the algorithms we looked into we decided to use: Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Naive Bayes, Decision Tree, K-Nearest Neighbor, Neural Network, Support Vector Machine, and Random Forest. From here we have decided to look into our dataset itself, to clean up and prepare for usage and what kind of visualizations we wanted to use for what aspects of the data. Analyzing with our own outcomes as looking into the findings of what others have found would lead us to better understanding our data and its outcome.

# 3 Approach

Before beginning our analysis, we must first inspect the raw data to see what prepossessing steps are necessary to have a fully cleaned dataset. The data contains 14 features and the target variable, income, which is a string divided into two classes: $\leq 50k$ and $\geq 50k$. Our dataset contains 48,842 data points, and about 7% (3,419) have missing values. We decided the best approach was to delete the rows with missing values since there was no good way to fill in all the missing categorical variables. We next modified our Income variable by translating all values of $\leq 50k$ and $\geq 50k$ into 0 and 1 respectively to make it easier to run our classification models later on. We removed 3 columns of data that we considered unneeded, that being the 'education', 'fnlwgt', and 'relationship' columns. We removed the education feature since the 'education-num' feature provided the same information in a numerical format. The integer value of educational levels are:

| | | | |
|---|---|---|---|
| Preschool | 1 | HS-Graduate | 9 |
| 1-4th Grade | 2 | Some College Experience | 10 |
| 5-6th Grade | 3 | Associate-Vocational | 11 |
| 7-8th grade | 4 | Associate-Academic | 12 |
| 9th | 5 | Bachelors | 13 |
| 10th | 6 | Masters | 14 |
| 11th | 7 | Professional-School | 15 |
| 12th | 8 | Doctorate | 16 |

The column 'fnlwgt' stansd for final weight, which is essentially the number of people the census believes it represents. This feature would be difficult for the model to interpret since it does not have a direct correlation with the income level. The 'relationship' column is removed because it highly correlated to the other column 'marital-status' which is a more personnel descriptor of the individual data point rather than its relationship to others that the 'relationship' column represents. We have also redesigned the 'native-country' data column to store the countries to their representative continents to prevent less skewed outcomes based on this data point, as well as making the

United States its own category in the 'native-country' column due to its large presence in the dataset. We wanted to keep this in order to analyze what different data points combine lead to a larger annual salary in different countries.

To tackle the many categorical columns in our dataset, we use the feature engineering tool: One Hot Encoder function, to change our categorical data to be more easily read by our chosen machine learning functions. Those columns that we need to change are: ['workclass', 'marital-status', 'occupation', 'race',"native-country"]. The default headers assigned by the encoder function were vague, so we created some code to replace the default labeled columns to have more understandable names.

Since this project is a binary classification problem, we chose nine different models to analyze our data. Each of these models has its advantages and disadvantages, which are presented below.

### 3.1 K Nearest Neighbors

- Pros: These algorithms can be used for classification, ranking, regression (using neighbors average or weighted average), recommendations, missing value imputation etc.
- Cons: It is a distance based-approach hence the model can be badly affected by outliers, it's prone to overfitting.

### 3.2 Logistic Regression

- Pros: Easy to separate response into 0 and 1 indicating $\leq 50k$ and $\geq 50k$.
- Cons: It can overfit in high dimensional datasets and does not support non-linear relationship between the predictor and the outcome.

### 3.3 Decision Tree

- Pros: Easy to understand and interpret, perfect for visual representation.
- Cons: It is very sensitive . Small change in the data can affect prediction greatly (High variance). For the decision tree methods, in addition to the unpruned approach, the prune, bagging and random forest approaches can be used to obtain better results.

### 3.4 Quadratic Discriminant Analysis

- Pros: Classification is usually more accurate and tends to outperform KNN and LDA.
- Cons: This model uses Gaussian assumption and complex matrix ops.

### 3.5 Linear Discriminant Analysis

- Pros: It is a simple, fast and portable algorithm.
- Cons: It requires normal distribution assumption on features/predictors.

### 3.6 Naive Bayes

- Pros: NB classifier performs better compare with other models like logistic regression and you need less training data.
- Cons: There is also the assumption of independence in predictors. In real life, it is almost impossible that we get predictors which are completely independent.

### 3.7 Neural Network

- Pros: Generally, more accurate than most models such as decision trees or KNN. They are able to capture and model non-linear relationships in data. Additionally they are highly flexible, and can be easily combined with other algorithms or techniques to create powerful machine learning models.
- Cons: It is very difficult to understand why it makes a certain prediction. They can be computationally intensive, which can make them slow to train and use. Not to mention they can be prone to overfitting, which means they may perform well on the training data but not generalize well to new data.

### 3.8 Random Forest

- Pros: Random Forests are resistant to overfitting, which means they can generalize well to new data. Random forests are highly interpretable, which means they can provide valuable insights into the relationships and patterns in your data. Not to mention they can easily train with most data, including larger sized ones.
- Cons: They can be difficult to optimize, and may require careful tuning of the hyperparameters to achieve good performance. They can be sensitive to the quality and quantity of the data, and may not perform well on imbalanced or noisy datasets.

### 3.9 Support Vector Machine

- Pros: SVM's are relatively memory efficient and work well when there's a clear margin of separation between classes.
- Cons: SVM's can be computationally intensive, which can make them slow to train and use, even more so than Neural Networks which ended up taking longer than 30 minutes to compute.

After cleaning our data, reorganizing it, and picking the machine learning models we wanted to use, the next step would then be to begin our test and train split of our data. We chose to use a 15% 85% test train split since this was giving us the best results with trial and error. We also ran cross validation on each of the models to further lower the variance caused by the random test train splitting. We applied this to three versions of our dataset. We used the original dataset after prepossessing as the first version. We then added a standard scaled and normalized version as well. The idea behind this was further scaling the data could be useful for columns such as the capital gain and loss since they are scaled much differently than age.

We tested each of our models individually, tuning each of the hyperparameters till we optimized their performance. We then put these optimized models into a pipeline to streamline the process of recording the results and plotting the accompanying visualizations.

## 4 Evaluation

There were three main metrics that we collected for each of the models we tested. These metrics were: accuracy, F1 score, and AUC. Of these metrics we most closely looked at the AUC value since it provides more information on the models performance. AUC is affected by the True Positive Rate and False Positive Rate of the model across different cut-off thresholds. In conducting a ROC curve analysis the Logistic Regression model had the highest AUC and Accuracy compared to other models. Hours-Per-Week and Age had the highest influence on the binary classification of $\leq 50k$ and $\geq 50k$. We also produced different visualizations of our data points and model performances. Created some Confusion Matrices to evaluate our work which did support the performance of Logistic Regression and also showed QDA having many false Positives, which indicated not great performance. There was also the inclusion of Correlation Matrix on our data before and after the One Hot Encoding to look at the relations between the different columns. Before the encoding we saw a strong relation to education-num and income. After the encoding there were some more strong relations with age, gender, and even a specific marital-status of civilian-spouse. There was the development of Violin Plots between income and some of the higher valued data points, such as age, hours-per-week, and educational-num. The income and age plot produced a nicer and smoother outcome.

## 5 Related Work

Since this data came from kaggle, there are multiple notebooks that analyze the data with machine learning models. Ultimately many people have their own opinions on what models to pick, what visualizations to show, and what data points to focus on. While we took their work into consideration we decided on picking the most well-known and accurate models to work with our data. We wanted to focus on models and graphs that we covered in class. Our methods revolved on what was taught as being the most useful for categorical datasets. When running into issues with our code we looked up possible solutions from multiple sources, but ultimately there wasn't a main source of work that we based our own approach off of. There was interest in using a pipeline method to run our many models, a website written by Preeti R. on the Top Machine Learning algorithms to use[1], and mentioned the usage of pipeline method to use with our models. This was the only form related work that was referenced when working on parts of our code.

## 6 Conclusion

Classification supervised learning gave us the ability to locate multiple trends within the data set to create recommendations of which variables produce significance. We initially decided to look at a handful of different analytical models for our data set, that being Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Naive Bayes, Decision Tree, K-Nearest Neighbor, Neural Network, Support Vector Machine, and Random Forest. We found the Quadratic Discriminant Analysis to be the least accurate function for our datasets. It performed the lowest scored for both accuracy and AUC. Through cross validation we were able to assure that our models weren't overfitting or underfitting the dataset. Ultimately after comparing all these functions we found that Logistic Regression produces the most accurate results for our dataset.

Through the power of our machine learning algorithms the dataset was capable of being able to predict individuals who would make $\geq 50k$ annual income. We saw how certain variables such as hours-per-week were a strong point associated with whether a person would make more or less than 50K annually. Age was also a very strong in the list of highest variable associated with making more than 50k annually. From these results our group can suggest which careers are lucrative and the variables associated with these individuals. Our results can also give us a better understanding of which professions and associated variables were under-performing in the market.

## 7 Citation

[1] Preeti R. "Top Machine Learning Algorithms for Classification." Towards Data Science,
URL: https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501