Exploring Classification Algorithms on a Census Income Dataset

Yuwen Yu
Alexander Castro
Brendan McShane
Ismail Mustafa
CISC 5790: Data Mining

Table of Contents

`

1 Introduction

This report is an analysis of data from the UC Irvine Machine Learning Repository consisting of entries of income data from the census bureau database. The goal is to explore different learning algorithms to predict whether a person makes over $50,000 a year. A test file consisting of data without income entries was provided and used for evaluating algorithm performance.

First, exploratory analysis of the data is performed in order to determine the necessary pre-processing tasks to prepare the data for analysis.

2 Pre-Processing

2.1 Exploratory Analysis

Initial exploratory analysis of the data showed that there are 32,561 entries each with the following features:

- **age** (continuous)
- **fnlwgt** (continuous)
- **capital-gain** (continuous)
- **capital-loss** (continuous)
- **hours-per-week** (continuous)
- **education-num** (continuous)
- **relationship** (categorical)
- **race** (categorical)
- **sex** (categorical)
- **native-country** (categorical)
- **marital-status** (categorical)
- **occupation** (categorical)
- **education** (categorical)
- **workclass** (categorical

In the following figures we explore the relationship between features in our dataset, as well as their relationship to the feature being predicted, *income*.

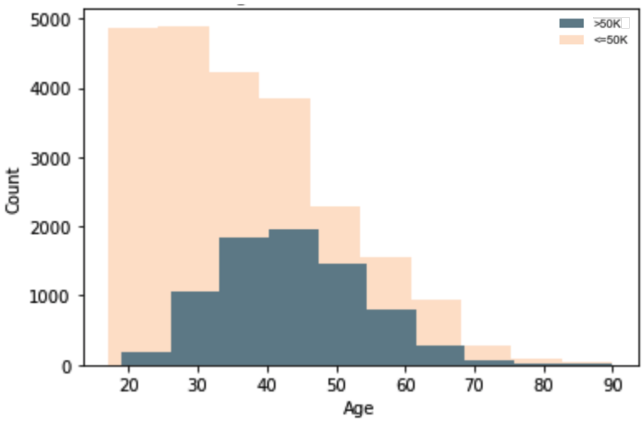


Figure 1 - Histogram of Age by Income Group

Based on Figure 1, it is clear that in every age category, more people make less than or equal to $50,000 annually than those who make more than

$50,000 annually. However, it is important to note the disparities in different age groups. For example, the gap in population difference based on income is much higher when people are in their 20s and 30s. However, this difference begins to narrow as people reach their midcareer in their 30s and 40s. The proportion of people who make more than $50,000 is much higher once people reach their 40s and above, with respect to their age category. This may be attributable to the fact that younger people tend to have entry-level jobs that pay much less as opposed to those who have much more experience and a higher salary due to having many years in the professional sector, which is a key result of age.

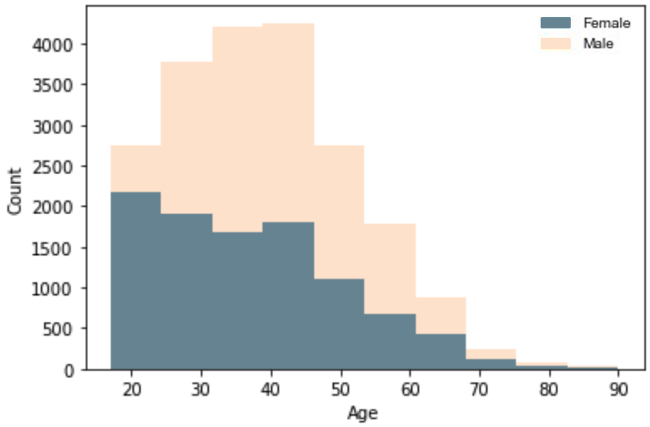


Figure 2 - Histogram of Age by Gender Group

Based on Figure 2, it is clear that in every age category, there are more males in the population than females. However, it is important to note the disparities in different age groups. For example, the gap in population difference based on gender is much higher when people are in their 30s and 40s. However, this difference is much narrower with people in their 20s as well as later ages such as 60+. This may show two key points in that females tend to live longer than men and that, proportionately, more females are being born into the population than males which may shift demographics in the future.

`

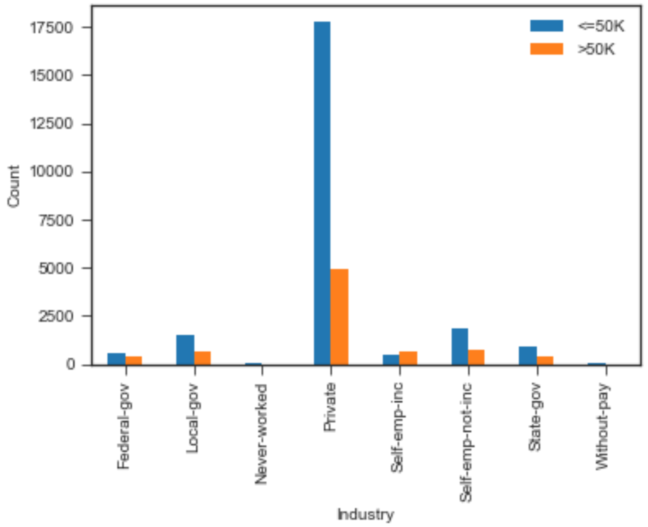


Figure 3 - Income by Industry

Based on Figure 3, it is clear that a majority of the population works in the private sector as more than half the population accounts for this industry. In addition, it is also observed that in most industries, a majority of the people are making less than $50,000 annually, with the exception of "Self-empl-inc" which represents incorporated self-employers. The breakdown of different categories may be better if they were placed into more broad categories such as "Government", "Private", "Self-Employed", and "Not Employed." This will allow for much more data points in each category and allow people to see more clearly which broad industry may lead to a higher pay. This will also prevent bias in the data as a lot of the data points currently come from the "Private" industry.

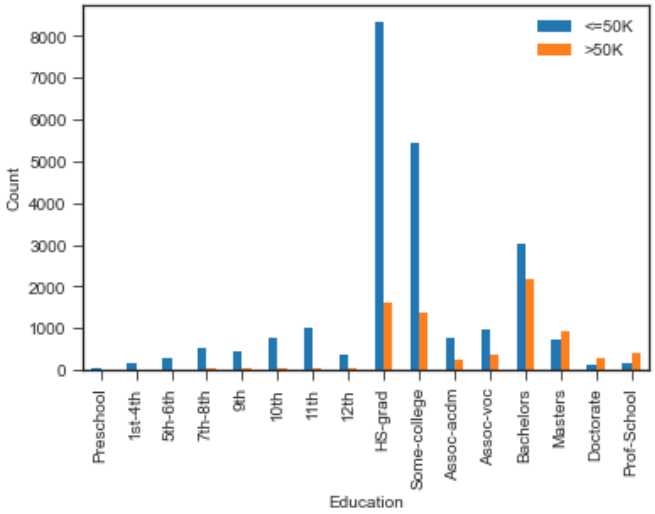


Figure 4 - Income Level with years of Education

Based on Figure 4, it is clear that a majority of the population have graduated high school or have even pursued higher education. In addition, it is also observed that in most education levels, a majority of the people are making less than $50,000 annually, with the exception of graduate-level education such as "Masters", "Doctorate", and "Prof-School". The gap between the difference in population's income disparity is much smaller once people graduate high school as less people are proportionately making less than $50,000 annually. This may show that pursuing higher education may lead to getting a higher salary. The breakdown of different categories may be better if they were placed into more broad categories such as "Grade School", "College", and "Graduate". This will allow for much more data points in each category and allow people to see more clearly which broad education level may lead to a higher pay. This will also prevent bias in the data as a lot of the data points currently come from the "HS-grad" educational level.

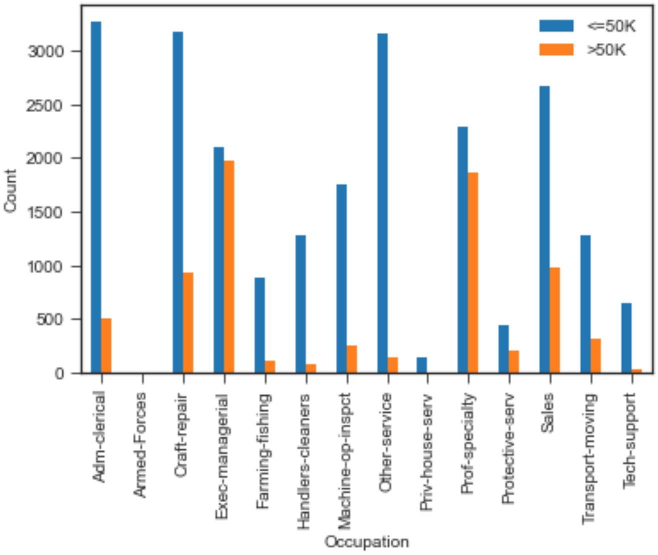


Figure 5 - Income by Occupation

Based on Figure 5, it is clear that there is no clear majority of the population attributable to a single occupation as the data seems to be spread across the various occupations. In addition, it is also observed that in every occupation, a majority of the people are making less than $50,000 annually. The gap between the difference in population's income disparity is much smaller in occupations such as "Exec-managerial" and "Prof-specialty" are proportionately making less than $50,000 annually. This may be attributable to the fact that these positions are more "white-collar" professions that require more experience.

`

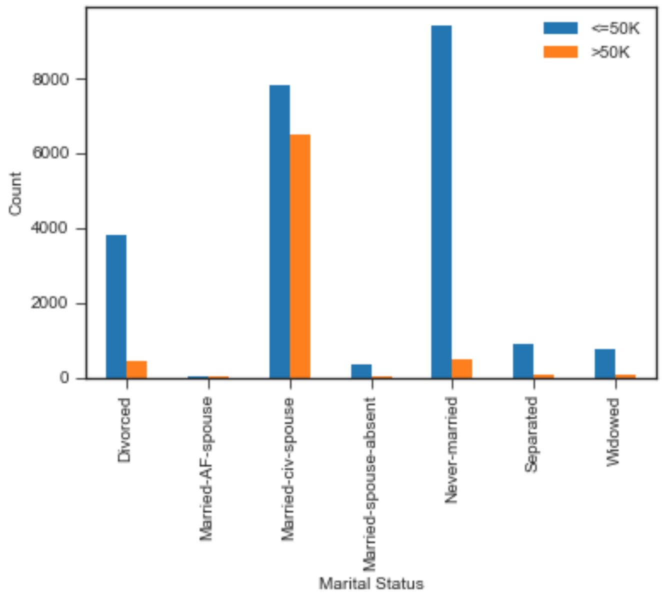


Figure 6 - Income by Marital Status

Based on Figure 6, it is clear that a majority of the population fall under the categories of “Married-civ-spouse”, “Never-married”, and “Divorced”. In addition, it is also observed that in every marital category, a majority of the people are making less than $50,000 annually. The gap between the difference in population’s income disparity is much smaller in the “Married-civ-spouse” category. This may be attributable to the fact that people who are married are usually older and established with financial security as opposed to people who have never married where the gap is much larger as they are still young and looking for low paying entry-level positions to start off their career. The breakdown of different categories may be better if they were placed into more broad categories such as “Married”, “Separated”, and “Never Married”. This will allow for much more data points in each category and allow people to see more clearly which broad marital status may lead to a higher pay. This will also prevent bias in the data as a lot of the data points currently come from only the two main categories.

`

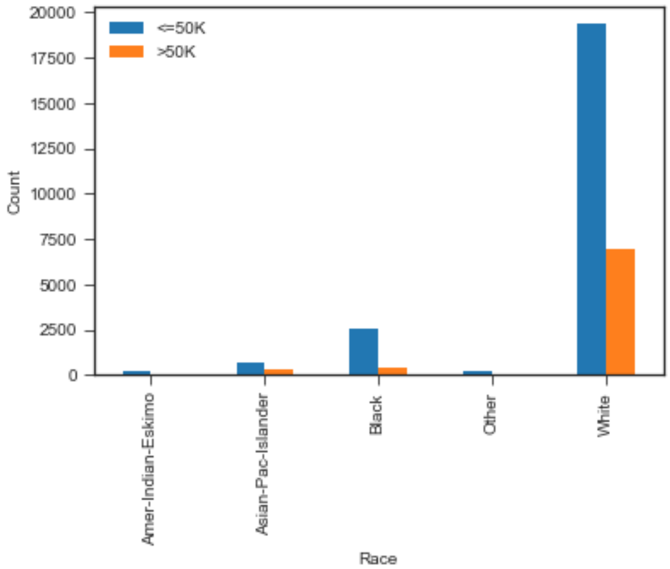


Figure 7 - Income by Race

Based on Figure 7, it is clear that a majority of the population fall under the race category of "White". In addition, it is also observed that in every marital category, a majority of the people are making less than $50,000 annually. The gap between the difference in population's income disparity is much smaller in the "White" category. This may be attributable to socioeconomic factors that provide inequalities where white people are more well off than minorities and people of color. It is important to note when evaluating the data and running tests, that the data may be much more biased towards the "White" population as almost 75% of the data comes from this category. It will be key to offset these biases to create a fair model.

2.2 Handling empty data

Analysis of the data showed that the data consisted of entries which were empty. The following methodology was used for determining the appropriate way to handle empty data:

1. Determine the percent of the data in the column that is missing
2. If the column has more than 6% missing data, impute the value
3. If it is less than 6%, delete the rows where the missing values lie

Missing values exist in the dataset for the workclass, occupation, and native_country fields of the data. Missing values make up 5.63%, 5.66%, and 1.79% of the workclass, occupation, and native_country fields, respectively, falling within the threshold to delete the rows containing these missing values.

`

2.3 Feature Selection

In the feature selection task, a filtering method was used to determine which columns were to be used for the model. The correlation criteria method allowed for the creation of a matrix demonstrating the most correlated feature to the class label.

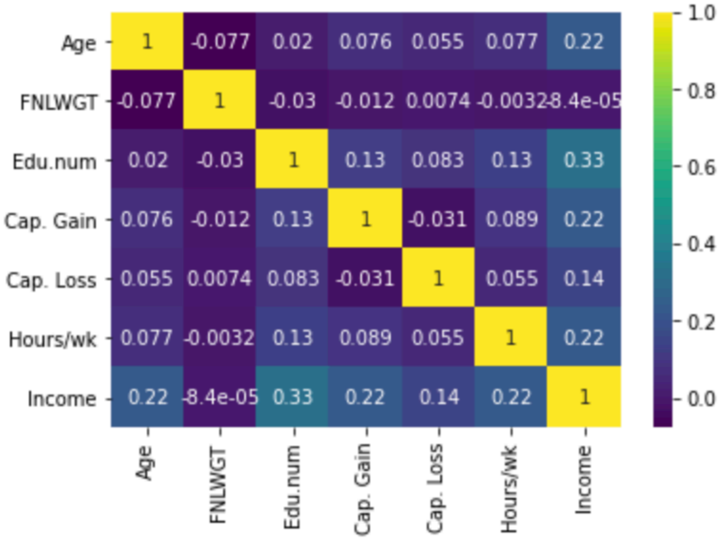


Figure 8 - Training Dataset Pearson Correlation Matrix

Below is an outline of the work we did in preparing each feature for our models:

1. **Age**: Separated into 4 bins: 0-20, 20-40, 40-60, and 60+. (the maximum value in this column was 90 years old)
2. **FNLWGT**: This metric measuring socio-economic demographics had little correlation with the target variable along with the rest of the features.
3. **Hours/wk:** In the US part time work is considered to be 20 hours or less, while full time work is around 40. This was the reasoning behind the separation into the bins 0-20,21-45, and 46+ hours per week.
4. **Capital Delta:** The Capital Gain and Capital Loss are measuring the same value. Capital Delta was created using the formula: Cap Gains - Cap. Losses. This feature also remains a continuous variable.
5. **Native Country:** 91 percent of ‘Native Country’ was the US, so the feature was made into two categories: ‘US’ and ‘Not Native to US’.

`

6. **Relationship:** Sex and Age may better reflect qualities that the relationship feature is attempting to measure. Also, it is a vague category; therefore, we decided to drop it from our models.
7. **Marital_Status:** For this column, ‘Divorced’, ’separated’, ‘Married-spouse-absent’ and ‘Widowed’ were put into **‘Separated’**. The ‘Married-AF-spouse’ and 'Married-civ-spouse' classes were considered as **‘Married’**. All others were classified as ‘Never-married.
8. **Occupation**: There were many possible outcomes that occupation had and an unclear method to make the feature more concise. We felt that Workclass was a more encompassing feature to use for the model, so Occupation was dropped from the model.
9. **Race**: The vast majority of the census takers were white, so similar to Native Country feature, we combined the non-white instances are put into one class called ‘other’.
10. **Sex**: This feature was kept as is because it is already a binary.
11. **Education**: Two modifications were made to this feature. All classes that were less than graduating high school were consolidated into ‘Grade School’. Also, the two associate-level degrees were combined into one ‘Associates’ class.
12. **edu_num**: This continuous feature was dropped from the model because it was vague and better represented by the education feature.
13. **Workclass**: Addressing the type of work that each census taker has, we consolidated some of the fields that we felt were overlapping. So 'State-gov', 'Federal-gov', and 'Local-gov' were combined into the ‘gov’ class.

3 Model Selection

3.1 Imbalanced Dataset

Figures 9 and 10 below show the balance of data when split on the two Income attribute labels for both the training and test datasets provided.

Label	Count %
<=50K	75.1%
>50K	24.9%

Figure 9 – Training Data Balance

Label	Count %
<=50K	76.38%
>50K	23.62%

Figure 10 – Test Data Balance

The four possible solutions below were proposed to handle imbalance data:

1. Change the performance metric (Solution selected)
2. Oversample minority class
3. Undersample majority class
4. Generate synthetic samples

We decided to use confusion matrices instead of using accuracy alone to handle the imbalanced data issue since these other methods introduce a certain level of bias to data.

3.2 Categorical Variable Encoding

The Ordinal Encoder from sklearn.preprocessing package was used for encoding categorical variables that have order: "Age", "Edu", 'Hours/wk'. The OneHot Encoder was used for encoding categorical variables that do not have order: 'Workclass', 'Marital_Status', 'Race', 'Sex', 'Native Country'. Finally, LabelEncoder was used to encode the target variable: ‘Income’.

3.3 Standardized Features

We used the StandardScaler function to standardize the features for Logistic regression and KNN since standardized features can improve model performance.

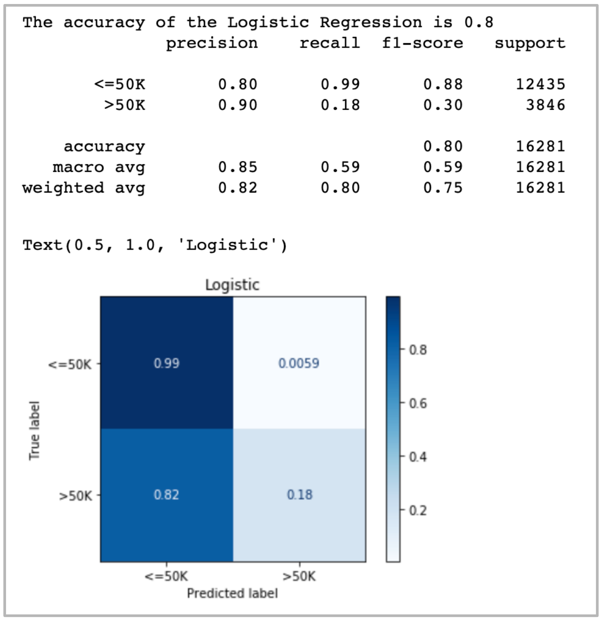
3.4 Build Model

Each of the following learning algorithms were tested in our prediction task. For most models, we have used the standard features included in the scikit-learn package with little modification.

- Logistic Regression with L2 regularization
- Random Forest Classifier
- K Nearest Neighbors Classifier
- Gaussian Naive Bayesian
- Ensemble model: Implement voting classification on the above four models

The results of running each in model in Python are outlined in the following figures.

`



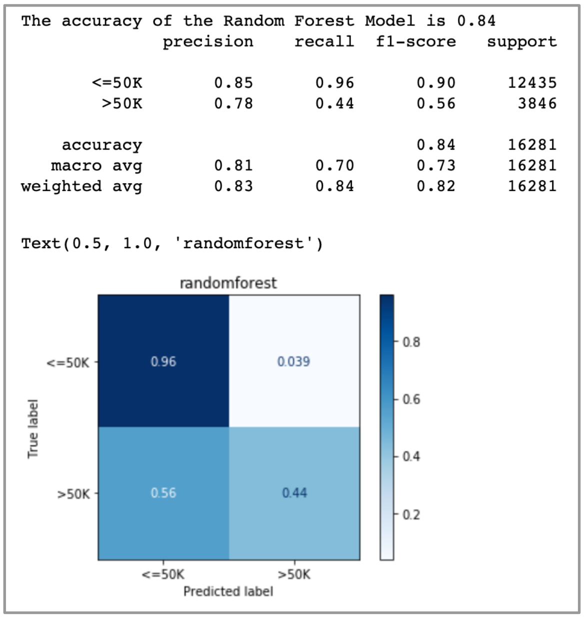
```
The accuracy of the Logistic Regression is 0.8
              precision    recall  f1-score   support

       <=50K       0.80      0.99      0.88     12435
        >50K       0.90      0.18      0.30      3846

    accuracy                           0.80     16281
   macro avg       0.85      0.59      0.59     16281
weighted avg       0.82      0.80      0.75     16281


Text(0.5, 1.0, 'Logistic')
```

Figure 11 - Logistic Regression Results



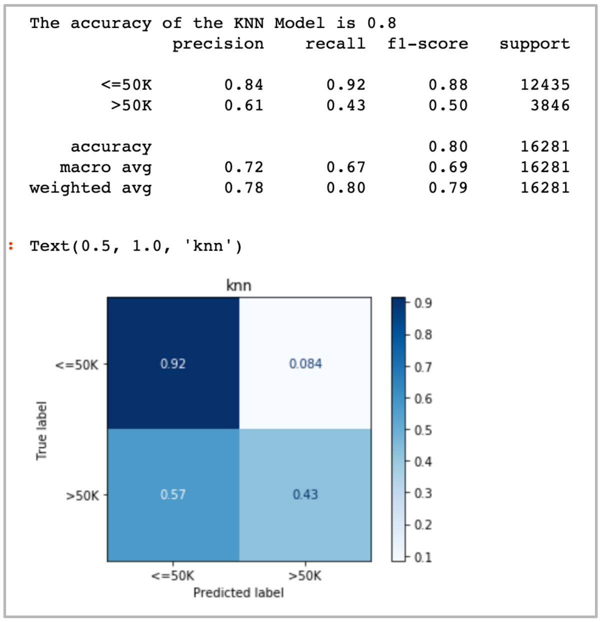
```
The accuracy of the Random Forest Model is 0.84
              precision    recall  f1-score   support

       <=50K       0.85      0.96      0.90     12435
        >50K       0.78      0.44      0.56      3846

    accuracy                           0.84     16281
   macro avg       0.81      0.70      0.73     16281
weighted avg       0.83      0.84      0.82     16281


Text(0.5, 1.0, 'randomforest')
```

Figure 12 - Random Forest Model Results



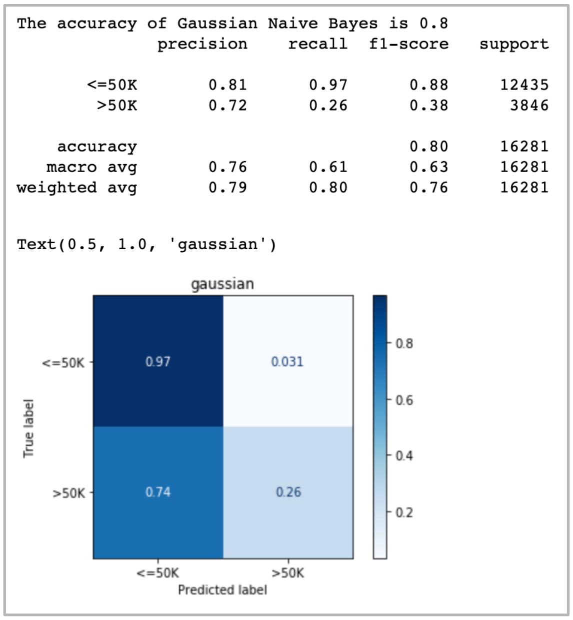
```
  The accuracy of the KNN Model is 0.8
              precision    recall  f1-score   support

       <=50K       0.84      0.92      0.88     12435
        >50K       0.61      0.43      0.50      3846

    accuracy                           0.80     16281
   macro avg       0.72      0.67      0.69     16281
weighted avg       0.78      0.80      0.79     16281


: Text(0.5, 1.0, 'knn')
```

Figure 13 - KNN Model Results



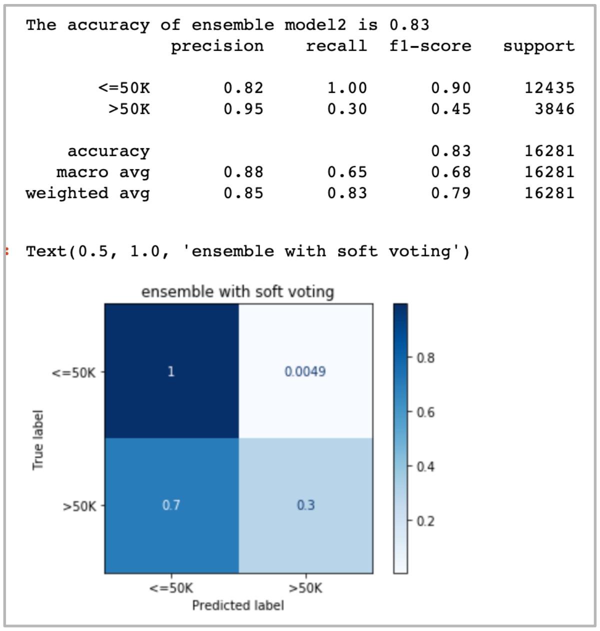
```
The accuracy of Gaussian Naive Bayes is 0.8
              precision    recall  f1-score   support

       <=50K       0.81      0.97      0.88     12435
        >50K       0.72      0.26      0.38      3846

    accuracy                           0.80     16281
   macro avg       0.76      0.61      0.63     16281
weighted avg       0.79      0.80      0.76     16281


Text(0.5, 1.0, 'gaussian')
```

Figure 14 - Gaussian Naive Bayes Results



```
The accuracy of ensemble model2 is 0.83
              precision    recall  f1-score   support

       <=50K       0.82      1.00      0.90     12435
        >50K       0.95      0.30      0.45      3846

    accuracy                           0.83     16281
   macro avg       0.88      0.65      0.68     16281
weighted avg       0.85      0.83      0.79     16281

Text(0.5, 1.0, 'ensemble with soft voting')
```

Figure 15 - Ensemble Model Results

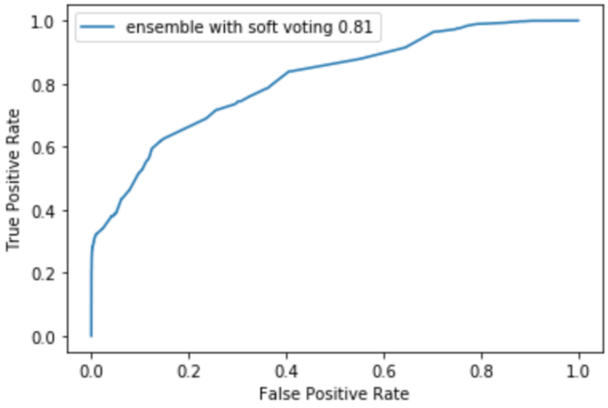


Figure 16 - Ensemble with Soft Voting AUC

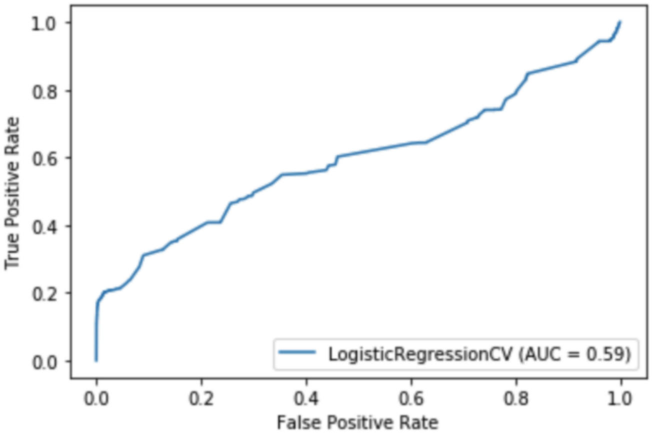


Figure 17 - Logistic Regression AUC

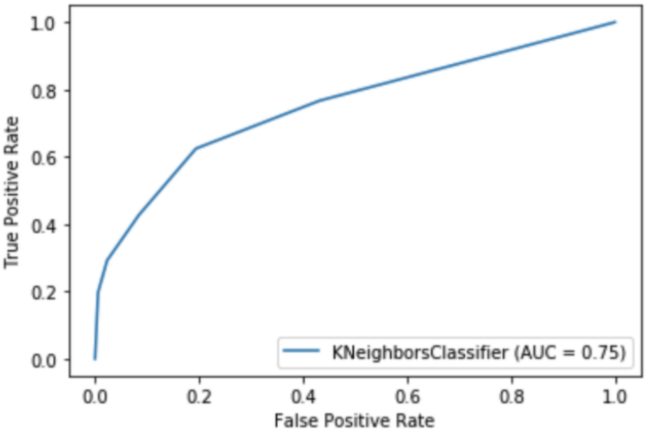
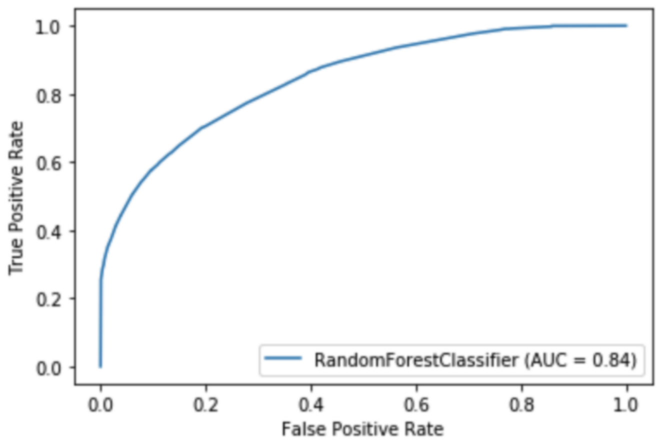


Figure 18 - KNN AUC

`



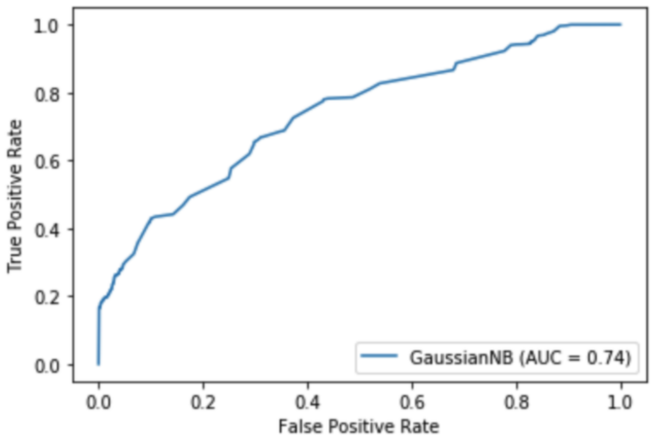


Figure - 19 Random Forest AUC

Figure 20 - Gaussian Naive Bayes AUC

4 Evaluation

In evaluating the models, accuracy may not be the best metric since the dataset is imbalanced. Therefore, we have evaluated our models mainly using ROC, precision, recall, and f1-score that can reflect true positive rate, true negative rate, false positive rate, false negative rate.

Precision measures how good our model is when the prediction is positive. Recall measures how good our model is at correctly predicting positive classes. In our case, we have f1 scores that are the weighted average of precision and recall.

Based on all the results from our models, Random Forest has the highest accuracy, f1 score, and ROC area.

From these recall and precision scores, we can see that the ensemble model has the highest recall scores on '<=50K' class and the random forest model has the highest recall score on '>50K'. Random forest performers better on capturing true '>50K' class that is a minority class in our dataset.

Therefore, we decided to pick the Random Forest model as it performs relatively well in this case. Moreover, random forest is able to handle missing value cases itself as well.

5 Conclusion

When examining the research as a whole, it is clear that the first building of the model produced important insights into the data, but there still exists a need for improvement in order to obtain results that represent the population with greater accuracy. One way to improve this research would be to work with a larger dataset. The data used for this research encompasses about 32,000 people which makes up only .01% of the entire population of the United States at the time, about 250 million. This is not

`

a significant portion of the population, and with more data points there would be more samples to test on which would better represent the population. In addition, another way to improve the research would be to test with more advanced learning algorithms such as deep learning. Using tools like this will only improve the accuracy of the results as it is better equipped to process and handle the data. A third way to improve the findings of this research would be to perform an imputation of the missing values. For the purposes of this initial research, the missing data was deleted due to the fact that it was less than the six percent threshold which we determined would signify that deleting the data would not have any significant impact on the findings. However, it would still be ideal to implement an algorithm such as KNN to impute the missing values as more data points will allow for a better representation of the population as a whole. Furthermore, this research should be tested again multiple times by separate parties, as this will allow for less bias and more verification of the findings.

It is important to consider when evaluating models that it is not possible to maximize both precision and recall because there exists a trade-off between them. When determining the best model, it is important to understand what the problem and goal we are trying to achieve are, as this will guide our selection. For example, consider the case where a credit loan company tries to use income to predict whether or not a borrower will default on a loan. If the company prefers to conservatively make decisions, they will want to be correct if the model detects '>50K' to reduce their risk exposure, so the random forest model would be a good fit. However, if the company wants to maximize profit or market share, they will want to be correct when the model detects '<=50K', so then the ensemble model would be a good fit since it has the highest recall and precision score.

6 References

[1] Becker, Barry, and Ronny Kohavi. "Census Income Data Set." *UC* Irvine *Machine Learning Repository*, 1 May 1996, archive.ics.uci.edu/ml/datasets/Census+Income.

`