20192032 Eldor Fozilov

IE30301-Final Report

Problem statement:

Predict whether an individual's annual income is above or below 50,000$ based on the following provided features.

Hypothesis:

Based on the given features, we might expect the following things below running any prediction model or analyzing the variable distributions in the dataset:

1. The average age of those who belong to above 50K class can be higher than that of those who earn less than or equal to 50K, and the age variable is going to be an important variable for income class prediction.

2. People who have higher level of education should earn more income than others, and the number of education years of individuals should significantly influence the probability of them earning more than 50K.

3. There can be more males in the above 50K category then females and the gender variables is going to be an important variable in predicting the income class.

4. The sector a person works in (in other words, working class) and what kind of job he or she does there should also greatly influence his or her income level.

(1) Exploratory data analysis

In the **data munging stage**, the important things that we did were the following:

1. We found out that variables such as “workclass”, “occupation”, and “country” contain “?” values, which we assumed them to be “NaN” values, and thus converted them accordingly.

2. We identified numerical and categorical variables, and saved their names in two lists, so that we can use utilize them later.

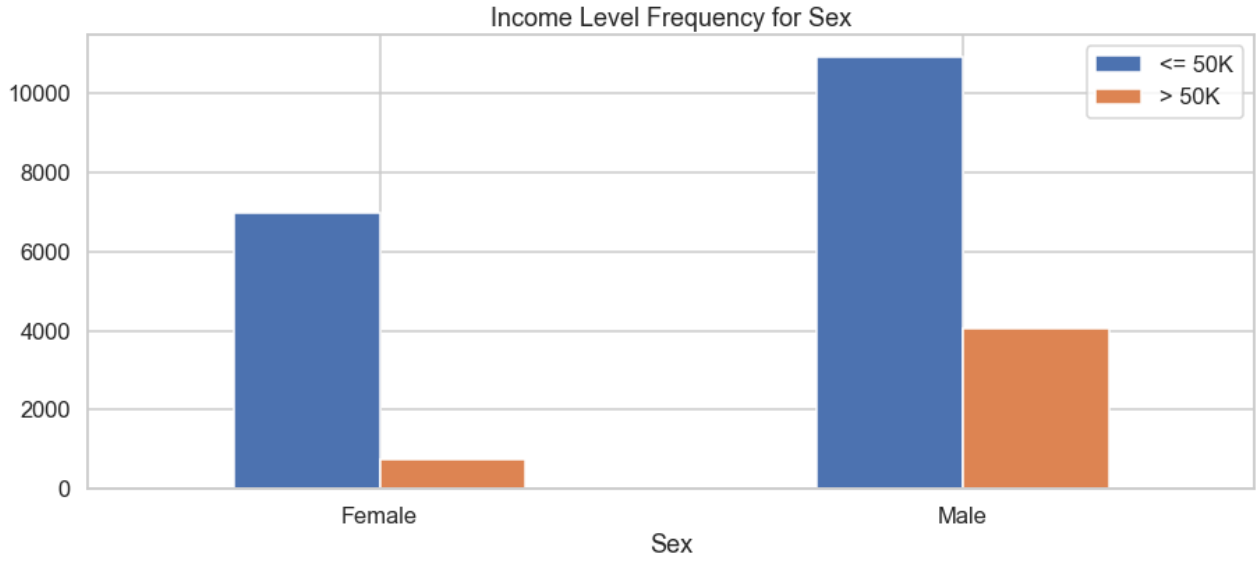
3. We found out that close 75% sample records in the dataset belong “<= 50K” income class, which might be an indication that out dataset is unbalanced.

4. After looking at the histogram of each numerical variable, we saw that variables such “capital gain” and “capital loss” have some significant outliers, and we removed those outliers so that they don’t disturb our analysis.

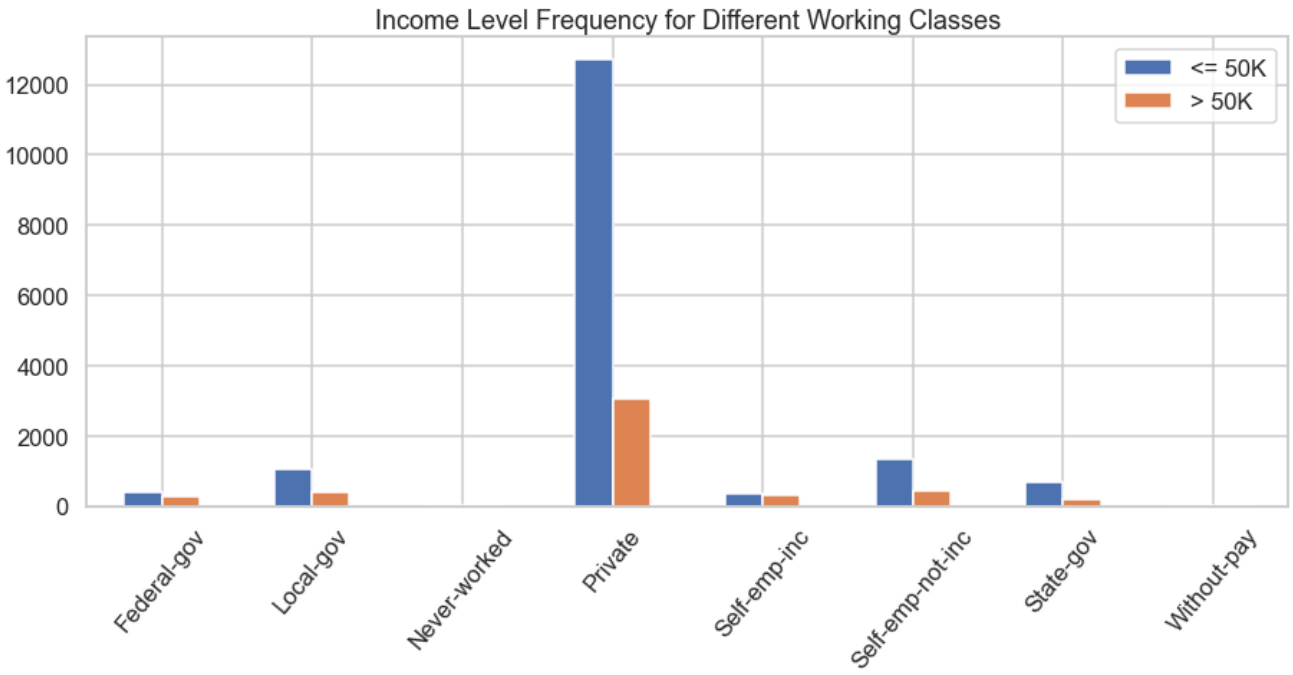
5. We used 75% of the data for training and 25% for training. Since the dataset is unbalanced, we decided to use stratified sampling method.

20192032 Eldor Fozilov

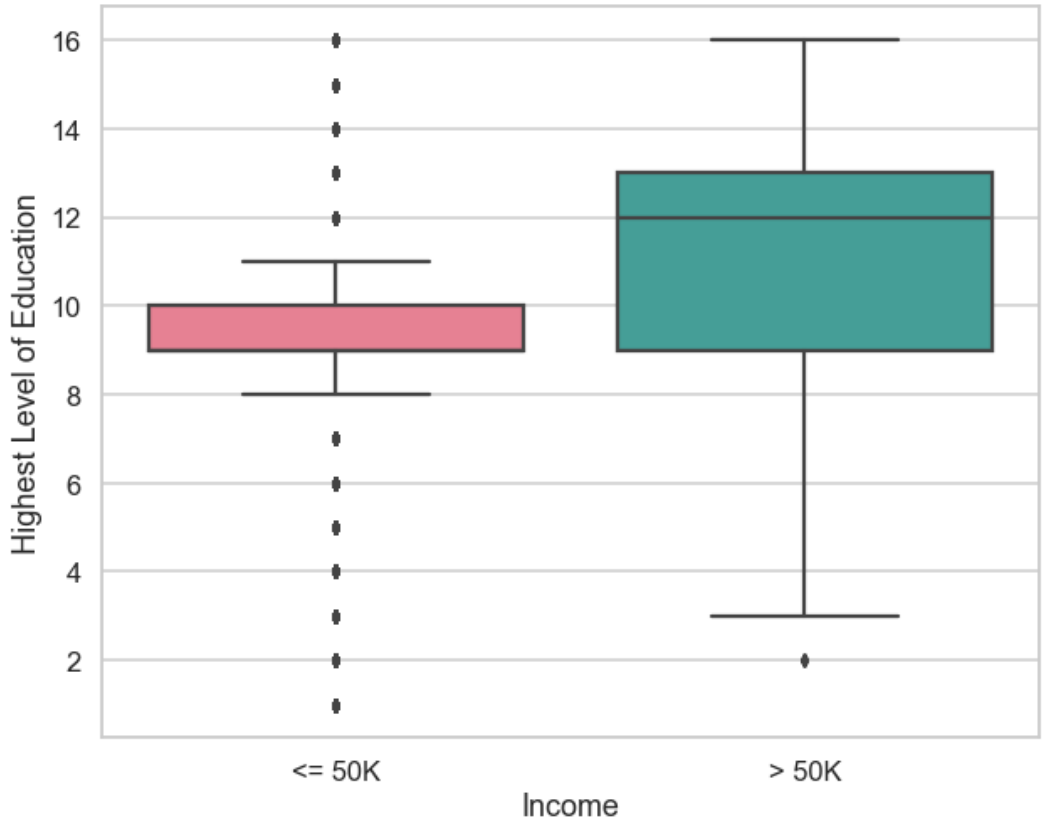
In the **EDA** stage, we found the following relationships between the income variable and feature variables:



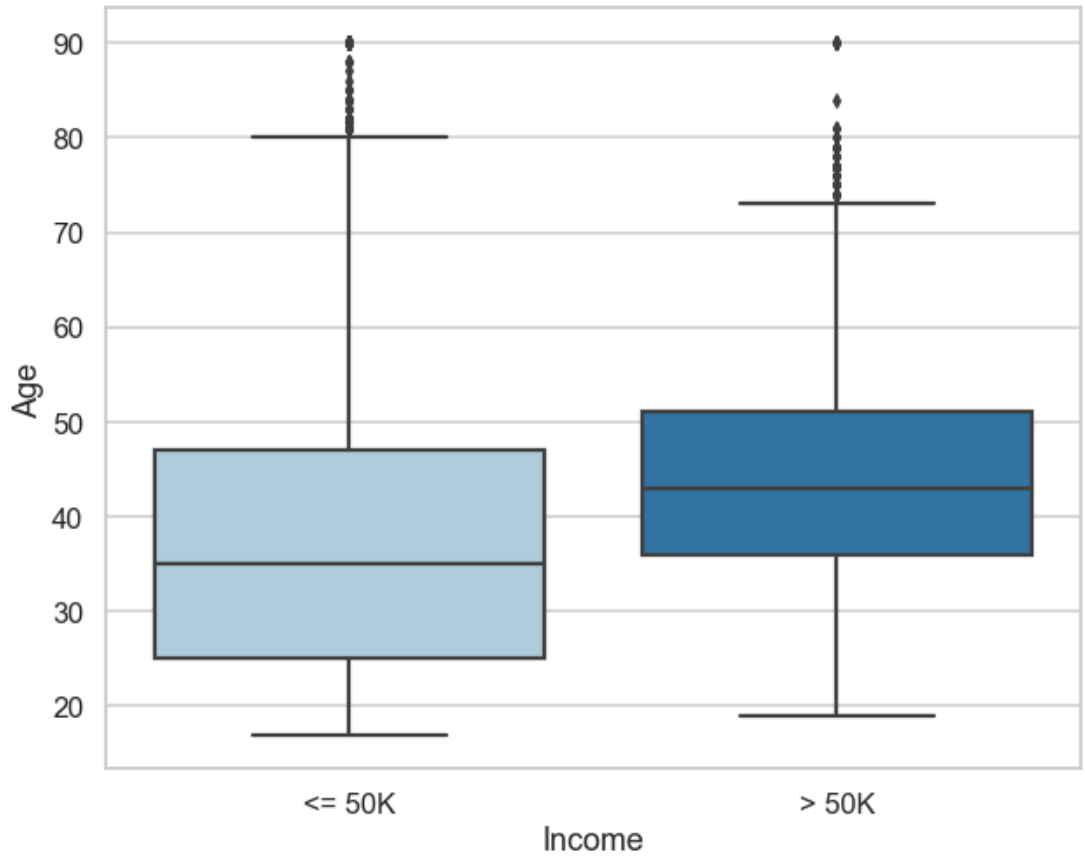
As we can see from the graph, males earn more income than females, which we expected to see.



Above, we can see that people working in the private sector earn the most amount of income compared to others and represent the majority of data samples in our dataset.



As we expected, the average number of acquired education years is higher in the above-50K income group.



The above graph also supports our hypothesis that the average age factor differs based on income class.

(2) Preprocessing

In the **preprocessing** stage, we did the following important things:

1. We decided not to consider the variable “education” because we thought that we have another similar (though numerical) variable “education_num”.

2. We filled the missing values with median for numerical variables and the mode for categorical variables. We also removed duplicate samples (but before that we split out the dependent variable as was recommended). We separately did the same operations on the test set.

3. We performed one-hot encoding for categorical variables and standard scaling for numerical variables. And after that, we combined the encoded categorical and scaled numerical variables into one dataset. We separately did the same operations on the test set.

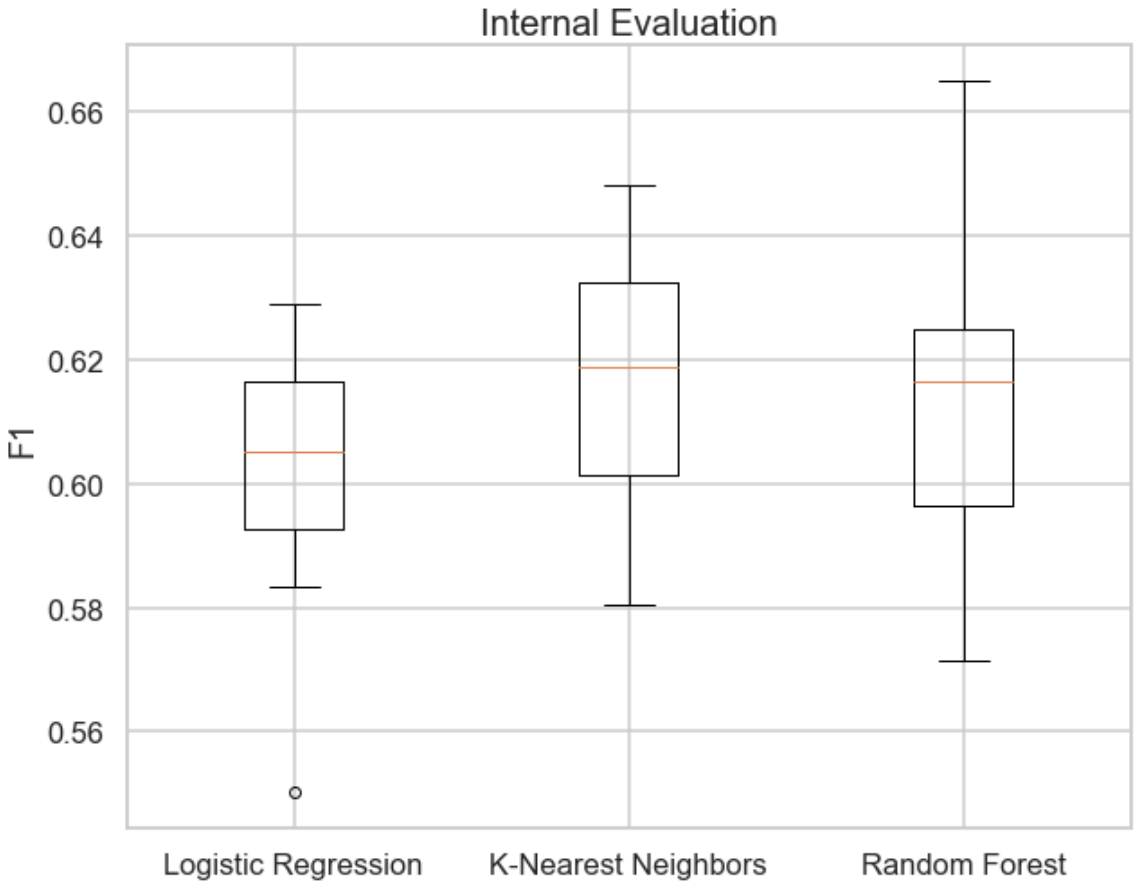
(3) Model train & test

In the model training and testing stage, we performed the following things:

1. We chose three models (KNN, Logistic Regression, Random Forest) to do the income classification.

2. We used RandomizedSearchCV to tune important hyperparameters of the above models.

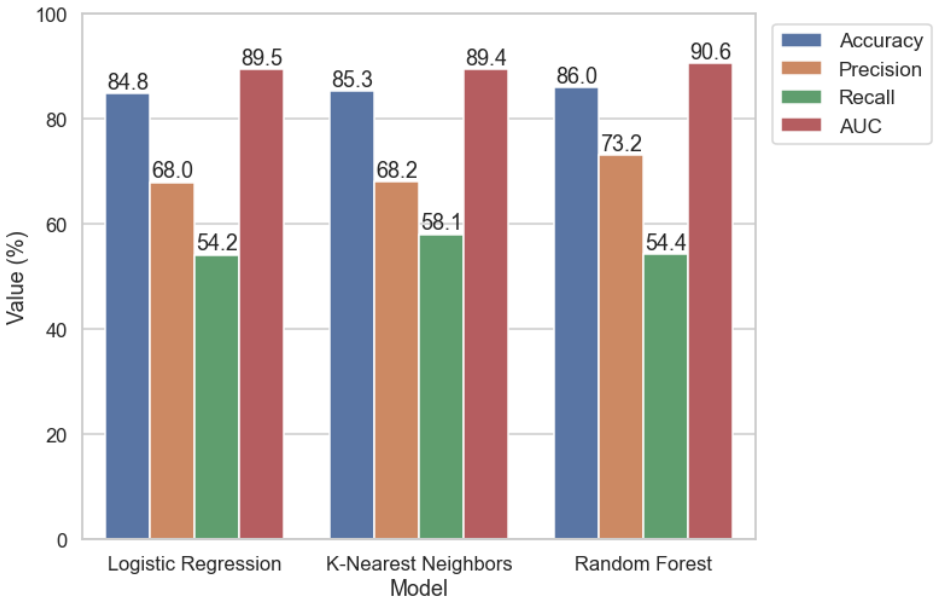
3. After finding optimal hyperparameters, we did an internal evaluation using Stratified 10 Fold method, where we chose F1 score as a performance measure since we are dealing with unbalanced and the accuracy is not the best measure. As a result, we go the following:



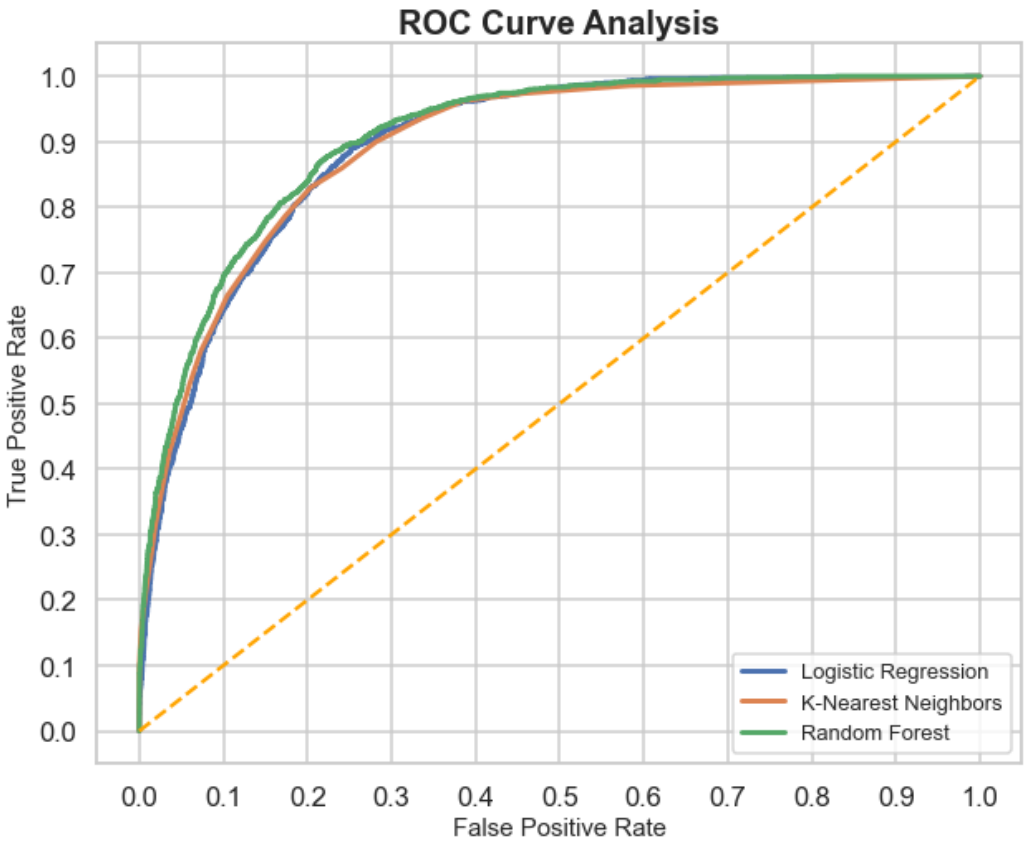
As we can see, KNN showed the best average performance. Random Forest also gave a almost the average result, however, it had a large variance. The lowest performance showed the Logistic Regression model.

(4) Result

Here are the results that we got by running the models on the test set:



We can see from the above plot, that all three models showed similar performance, Random Forest having a bit higher result than the other models in most evaluation metrics, except Recall, which was the highest in KNN model.



The same is true for the ROC curves of three used models. They are very close to each other. Though, Random Forest is a bit high to the upper left than other models (this was quantitatively shown in the previous bar plot).

(5) Discussion & Conclusion

To check the validity of our hypothesis, we did one more thing. We looked at the feature importance measure, which is a good quality of Random Forest model.

As we hypothesized, variables such as “education num”, “age”, and variables related to gender, sector, and occupation were among the top 20 most important variables in the Random Forest model, and which had a significant influence on the classification results.

This can be seen in the below graph.

20192032 Eldor Fozilov