Term Project Milestone 3 - Census Income Prediction

Team Name: Analytics

Team Members: Kesav Adithya Venkidusamy, Meenakshi Shankara, Sherry Kosmicki

Bellevue University

DSC630-T301 Predictive Analytics (2227-1)

 Professor Fadi Alsaleem

07/31/2022

**Census Income Prediction**

**Milestone 2 - Data Selection and Project Proposal (Week 2)**

### **Information about Dataset**

**Name**: Census Income

**About this dataset:**

This data was extracted from the [1994 Census bureau database](http://www.census.gov/en.html) by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). A set of clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0))

**Abstract:** Predict whether income exceeds $50K/year based on census data

|  |  |
| --- | --- |
| **Data Set Characteristics** | Multivariate |
| **Attribute Characteristics** | Categorical, Integer |
| **Associated Tasks** | Classification |
| **Number of Instances** | 48842 |
| **Number of Attributes** | 14 |
| **Missing Values** | Yes |
| **Area** | Social |
| **Date Donated** | 5/1/1996 |

**Attribute Information:**

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Feature Description** | **Feature Type** |
| age | Age of the person | Continuous |
| workclass | Work class of the person | Discrete |
| fnlwgt | Final Weight | Continuous |
| education | Education of the person | Discrete |
| education-num | Number of years the person | Continuous |
| marital-status | Marital status of the person | Discrete |
| occupation | Occupation of the person | Discrete |
| relationship | Relationship of the person to the family | Discrete |
| race | Race of the person | Discrete |
| sex | Sex of the person | Discrete |
| capital-gain | Capital Gain | Continuous |
| capital-loss | Capital Loss | Continuous |
| hours-per-week | Hour the person worked for a week | Continuous |
| native-country | Native country | Discrete |
| Income | Income of the person | Target |

### **What types of model or models do you plan to use and why?**

A logistic regression model will be used on the dataset to determine which features are mostly related or correlated to our target which is “Income” of the person. Logistic regression is a statistical analysis method used to predict a binary outcome such as yes or no based on prior observation of the data set. Here, “Income” feature present in the dataset has only binary values: whether the income of the person is less than or equal to 50K per year or greater than 50K per year. So, this feature will be used as target for the model. This model falls under supervised learning as the data is well labelled and has a target variable, a column in the data representing values to predict from other columns in the data. Sometimes supervised leaning is called predictive modeling. Supervised learning allows collecting data and product data output from previous experience.

Under supervised learning, this dataset falls under classification model as it reads the input and generates an output that classifies the input into two categories: one having income less than or equal to 50K per year and another with income greater than 50K per year.

In addition to the logistic regression model with 1 target and 14 features, another logistic model will be used with only 5 best features where the 5 features are selected based on highest chi-squared statistics.

### **How do you plan to evaluate your results?**

We plan to calculate the accuracy, precision, recall and F1 score of both the logistic regression models with 14 features and best 5 features. We will also verify the performance of the model by visualizing confusion matrix which is a 2\*2 table that shows the predicted values from the model vs. the actual values from the test dataset. We will plot ROC curve to determining the best cutoff value for predicting whether a new observation is a "failure" (0) or a "success" (1).

The Area Under the ROC curve (AUC) metric is evaluated to see how well a logistic regression model classifies positive and negative outcomes at all cutoffs. The value can range from 0.5 to 1. The result is considered excellent if AUC value is between 0.9-1, good for the AUC values between 0.8-0.9, fair for AUC values between 0.7-0.8 and poor for the AUC values 0.6-0.7 and failed for the AUC values between 0.5-0.6.

### **What do you hope to learn?**

Working as a group on this course project will help us to understand how to work with others in the same field of study. We can learn each other's strengths and highlight them in the process. For those that have areas they would like to improve, we can capitalize on learning from our teammates. Learning possibilities are available if we communicate our project progress.

During the review process of the income prediction dataset, we can determine if there are any statistical patterns or predictors based on visualizations. The modeling process can help us define the best model to use with our data based on accuracy, precision, and recall. Once we have trained our model new data can be applied to predict what income category a person is part of.

### **Access any risks with your proposal**

One of the earliest challenges we might face is during the data preparation step of the model building. Identifying the correct features that contribute to the target, planning on how to handle the missing values, deciding the next steps if the data is imbalanced to name a few. The way to mitigate these issues would be creating various visualizations to identify correlations. To mitigate data imbalance, we may choose to over-sample or under-sample the dataset. We may also need to go back to research other relevant supplement datasets to strengthen the cause.

During the model building phase, we may face the challenge of finding the right models for our project in terms of accuracy. As a mitigation plan, we will identify at least 3 models to train the data and calculate the accuracy with multiple methods.

We will also update/upgrade/change the course of our project based on the feedback received during peer reviews.

### **Identify a contingency plan if your original project plan does not work out.**

If for any unforeseen reasons, we are unable to continue working on this dataset, we have considered a backup dataset and we will perform a high-level analysis in parallel. This will help us quickly shift to the new data set without too much loss of time.

Backup Dataset - <https://www.kaggle.com/datasets/paradisejoy/top-hits-spotify-from-20002019>

### **Include anything else you believe is important.**

**Why income prediction is important**

Income prediction is important for a variety of areas in the private and nonprofit sectors. One critical area this affects is marketing, where income segmentation of the population is an extremely important tool. Businesses may make different variations of their items designated for certain subgroups of the population, and these subgroups often include the income of individuals. Income prediction also helps to identify those individuals who are of a lower income that may need the most assistance, who some nonprofits strive to identify and assist. The ability to predict the income of individuals from this information has far-reaching impacts for every industry.

**Data consideration for logistic regression**

Following are some of the important points to be considered while choosing data set for logistic regression.

* The response variable should be binary
* The features present in the dataset should be independent to one another
* Make sure the data represent the population of interest
* Collect enough data to provide necessary precision
* Measure variables as accurately and precisely as possible
* If the model does not fit the data, then the results can be misleading. In the output, use residual plots, diagnostic statistics for unusual observations, and model summary statistics to determine how well the model fits the data

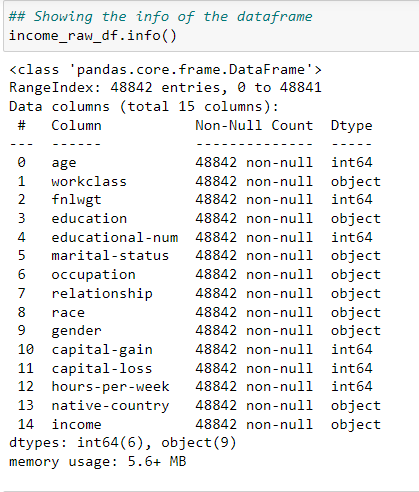
**Milestone 3 - Data Selection and Project Proposal (Week 6)**

### **Will I be able to answer the questions I want to answer with the data I have?**

The problem statement of this project is to identify the dataset feature(s) which are mostly related to or affecting the income of household. With dataset having total number of records as 48842, we would be able to predict or answer our problem statement. The dataset consists of 15 features of which 6 are numerical and rest all are categorical with “income” being the target. The target variable income contains 2 values <=50K and >50 which would be subsequently converted to 0 and 1 respectively. The details are shown in figure 1.

Among 14 features, we see missing/null values present only for below features. I have given the percentage of missing values for each of the feature in table 1. We noticed that the values where ‘workclass’ is missing, also has ‘occupation’ missing. While trying to identify the extra rows where ‘occupation’ is missing, we observed the workclass is ‘Never-Worked’. Since the percentage of null values present in these features is low, the rows will be removed from the dataset.

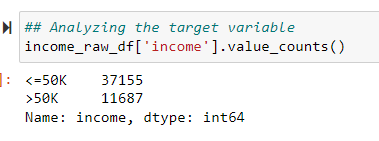
|  |  |  |
| --- | --- | --- |
| **Feature Name** | **# Of missing Values** | **Percentage** |
| workclass | 2799 | 5.7% |
| occupation | 2809 | 5.8% |
| native-country | 857 | 1.8% |



**Table 1: Features with null values and percentage**

**Figure 1: Features and dtypes**

The income column is our target variable with 2 values - ‘<=50K’ and ‘>50K’. The count of these values is 37155 and 11687 respectively, suggesting that people with income higher than 50K are significantly less, and our data set is kind of imbalanced considering the target variable. However, we will evaluate the outcome and apply filter to the dataset, if required.



### **What visualizations are especially useful for explaining my data?**

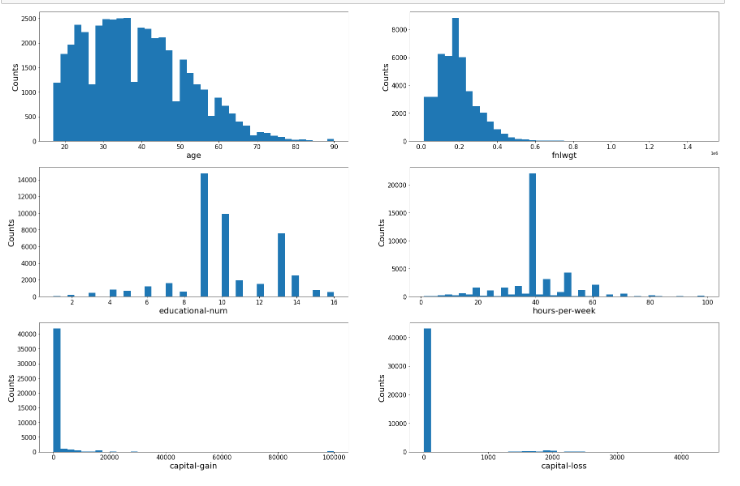
As mentioned before, the dataset contains 6 numerical features and 8 categorical features as follows, and ‘income’ feature being the target variable.

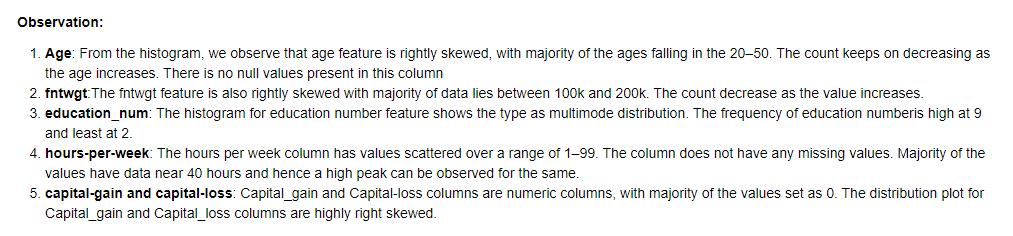
Numerical: age, fnlwgt, educational-num, hours-per-week, capital-gain, capital-loss

Categorical: workclass, education, marital-status, occupation, native-county, relationship, race, gender

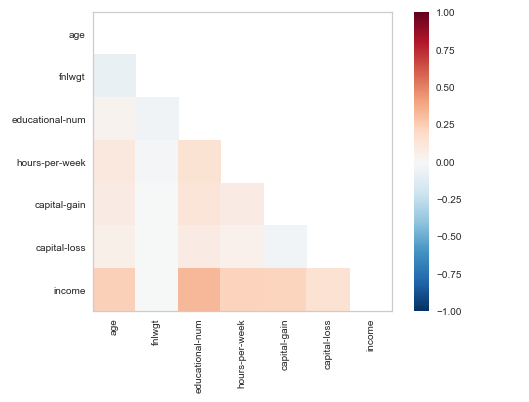
So, we have used several visualization charts to analyze the useful information present in the data. The following are the visualizations used based on nature of the feature.

***Histogram****:* Histogram is used to identify the distribution of numerical features present in the dataset.

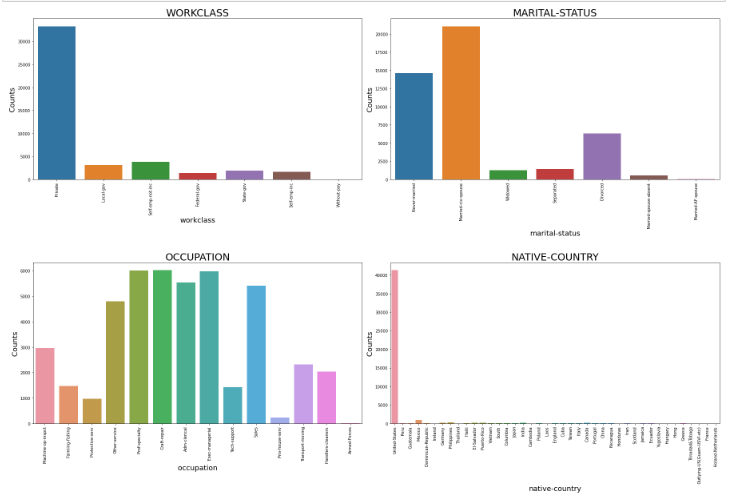


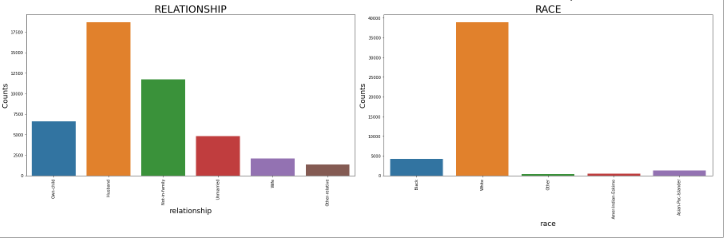


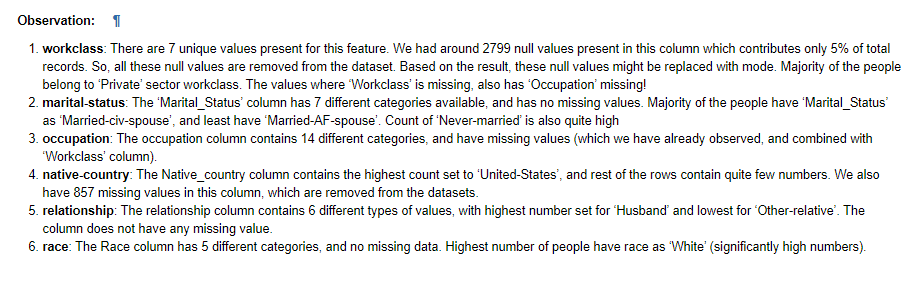
***Heat Map****:* Heat map has been created to understand Pearson’s correlation between the target variable ‘income’ and other numerical variables. We observed that all numerical features have positive correlation with the target except fnlwgt feature. Among the features having positive correlation, age and education-num features are having high value. The details are shown in the below heat map chart.

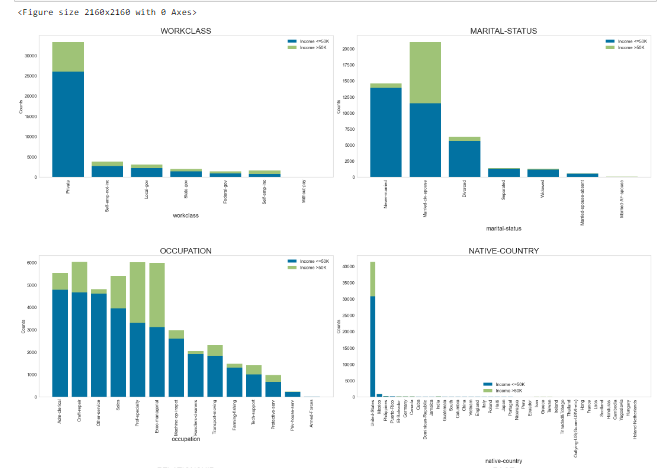


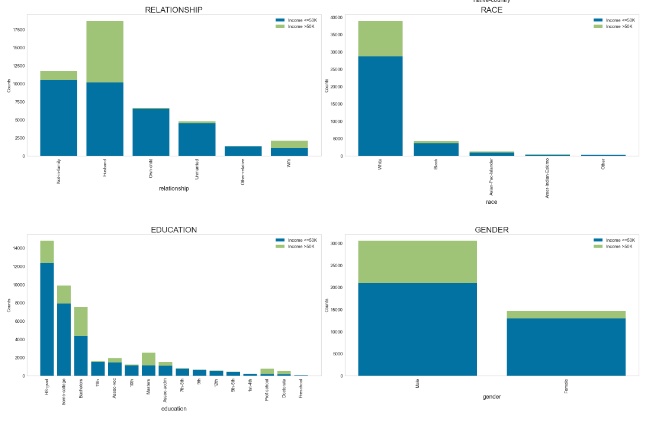
***Bar and Stacked Bar Graph:*** Bar graph has been plotted for all categorical features to understand the distribution of data among unique values. Stacked bar chart has been plotted to compared those earning less than or equal to 50K (represented as 0) and greater than 50K (represented as 1) for all the categorical features.

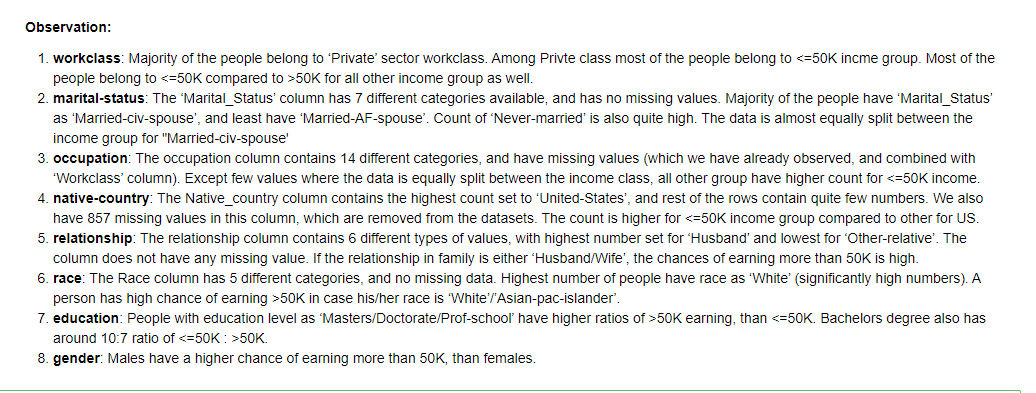












### **Do I need to adjust my model/evaluation choices?**

In Milestone 2 it was determined we would use logistic regression on the dataset. This analysis method is specific to a binary target variable which is our income determination amount. Once we’ve completed this step, calculating accuracy, precision, recall and F1 scores should help determine quantify model performance. The ROC curve metric will also be used to evaluate the logistic regression model where we hope to see values between 0.9-1.

The group has discussed creating a decision tree model to evaluate the data in addition to logistic regression. Based on categorical variables such as workclass and marital-status in this dataset, we can take them and group the values into a broad summarization of the values and use them for evaluating the decision tree.

The dataset will work well with this evaluation since a decision tree can handle both numerical and categorical values. The other benefits of using a decision tree are it works with the outliers and the missing values currently present in our dataset. We will have to determine the best way to split the variables into subgroups. This will help build efficiencies in the evaluation done by the decision tree. Additional data gathered for the purpose of continuing the evaluation of a person's potential income could strengthen the data model over time. We will then determine the accuracy of the decision tree and avoid an opportunity for overfitting.

### **Are my original expectations still reasonable?**

The original expectations are to find accuracy in the model building process to predict the income of an individual based on the gathered variables. The predictive data within the dataset has value and is functional based on the visual evaluations and data munging steps during the EDA process. If the accuracy of our models does not meet our needs, we may consider other model options such as Naïve Bayes classifier.

We will continue to meet once or twice weekly as a team to continue our discussion of the project. Each of us has a personal stake in the project. We will gather information and define the best possible steps to complete the project using the CRISP-DM process model.

**Milestone 4 – Finalizing Your Results (Week 8)**

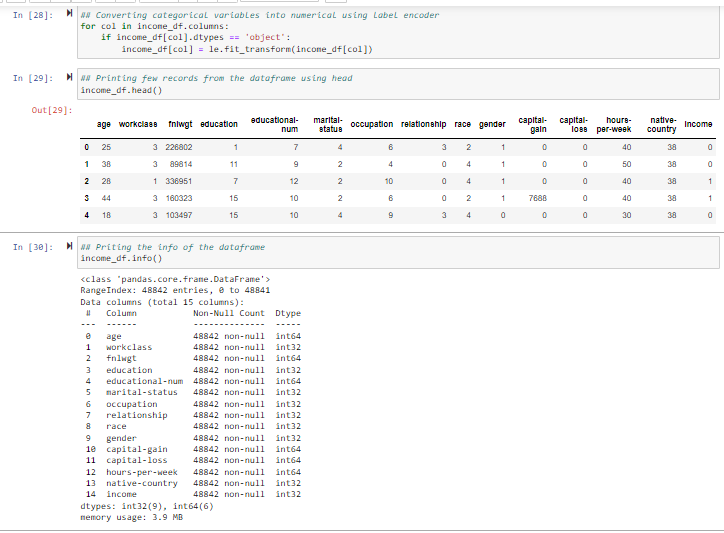
### **Explain your process for prepping the data**

As a team we worked individually prepping the data. This process created functional conversation among each of us to show the best possible data preparation method. We also found that each of us could bring a unique set of skills to ensure we covered all the necessary components when prepping the data.

While evaluating the data for null values, the workclass column has 2799 null values, occupation column has 2809 null values, and native-country column has 857 null values. The mode was calculated for each column and was used to replace the null values respectfully. A comparison was done to find duplicates within the data. A duplicate was decided using all columns for comparison. This evaluation determined 52 duplicates which were removed from the base dataset accordingly. After much discussion, we decided to use all the columns presented in the original data load.

As a team we discussed the rationale behind whether you split the data before or after choosing an encoding method. Team members tried both methods. We found supporting documentation that found the splitting of our dataset should be done first, and then the encoding method can be applied to the train data separately from the test data.

We started by using a label encoder method for the categorical variables. This method was then applied to our model choices and the accuracy measurements were evaluated. We also applied a standard scaler to our train data and applied the same model choices. Those results helped us decide the best encoding method.



*Figure: 4.1: Dataset after applying label encoder*

### **Build and Evaluate at least one model**

**Logistic Regression**

For this project, the first model we chose to build is Logistic Regression with all the features. The accuracy for this model was 83%. To further evaluate the accuracy, a confusion matrix was created. Although there was a high level of predicting incomes below 50K, the prediction on incomes above 50K was not enough as the false negatives seem to be higher.

Chart

Description automatically generated

Figure: 4.2 Confusion matrix for Logistic Regression

**Decision tree**

Next, we moved to the Decision tree model to see if this model improved the accuracy score of prediction. All the features were evaluated. The accuracy score for this model was came to 82% which is a slight reduction that the Logistic Regression model. From the confusion matrix for this model, we observe that the model did seem to be improve the predictions for income >50K but degraded in predicting incomes < 50k.

Chart

Description automatically generated

*Figure: 4.3 Confusion matrix for Decision Tree*

**Random Forest classifier**

The third model we decided to build is Random Forest classifier with all features. The accuracy score for this model is observed to be at about 87% which is the highest so far.

Chart

Description automatically generated

*Figure: 4.4 Confusion matrix for Random Forest Classifier*

We can observe from the confusion matrix that the predicted values for the target feature have greatly improved with 10470 predictions for income under 50K and 2212 correct predictions for income over 50K.

Please note that the above-mentioned models were evaluated considering various combinations of standardizing, feature removal. After multiple evaluations we have decided to consider all the 14 features and encode the dataset before splitting into train and test sets and then standardizing for predicting if the income would be under or over 50K.

### **Interpret your results**

As mentioned before, the following models have been applied on the data. I have mentioned the scores that we received for each of the model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **# of Features** | **Accuracy** | **F1 Score (Income <=50K)** | **F1 Score (Income >50K)** | **AUC Score** |
| Logistic Regression | All (14) | 79.80% | 0.88 | 0.40 | 0.74 |
| Decision Tree | All (14) | 82.14% | 0.88 | 0.62 | 0.76 |
| Random Forest | All (14) | 86.55% | 0.91 | 0.69 | 0.91 |
| Logistic Regression with StandardScalar | All (14) | 83.01% | 0.89 | 0.56 |  |
| Decision Tree with StandardScalar | All (14) | 82.17% | 0.88 | 0.63 |  |
| Random Forest with StandardScalar | All (14) | 86.54% | 0.91 | 0.69 |  |
| Logistic Regression | After removing unwanted fields (11) | 80.74% | 0.88 | 0.49 | 0.81 |
| Decision Tree | After removing unwanted fields (11) | 78.44% | 0.86 | 0.53 | 0.72 |
| Random Forest | After removing unwanted fields (11) | 82.32% | 0.89 | 0.60 | 0.87 |
| Logistic Regression with StandardScalar | After removing unwanted fields (11) | 81.07% | 0.88 | 0.50 |  |
| Decision Tree with StandardScalar | After removing unwanted fields (11) | 78.44% | 0.86 | 0.53 |  |
| Random Forest with StandardScalar | After removing unwanted fields (11) | 82.18% | 0.89 | 0.60 |  |
| Logistic Regression – Target variable with 5 best features using X2 | With Top 5 Feature | 79.99% | 0.88 | 0.48 | 0.81 |
| Decision Tree – Target variable with 5 best features using X2 | With Top 5 Feature | 79.56% | 0.87 | 0.54 | 0.77 |
| Random Forest – Target variable with 5 best features using X2 | With Top 5 Feature | 80.86% | 0.88 | 0.57 | 0.85 |

**Accuracy:** Accuracyrepresents the number of correctly classified data instance over the total number of data instances.

**F1 Score:** F1-Score is a metric which takes into account both precision and recall.

* **Precision:** Positive predictive value
* **Recall:** true positive rate

**AUC Score:** What area under the ROC curve describes good discrimination? We will use the following rule of thumb

* 0.5: This suggests no discrimination, so we might as well flip coin
* 0.5-0.7: We consider this as poor discrimination, not much better than a coin toss
* 0.7-0.8: Acceptable discrimination
* 0.8-0.9: Excellent discrimination
* >0.9: Outstanding discrimination

We have applied following 3 models on the dataset.

* Logistic Regression
* Decision Tree Classifier
* Random Forest Classifier

Initially, these models have been applied on the dataset having all the features. Then, we have removed below irrelevant features from the dataset and applied those modeling. Finally, we calculated the top 5 features present in the dataset using X2 and applied these modeling.

**Features removed from the dataset during 2nd iteration:**

* Capital-gain - Most of the rows are having value as 0
* Capital-loss - Most of the rows are having value as 0
* Fnlwgt - Not giving any meaningful information

**Logistic Regression:**

Initially, logistic regression has been applied of the dataset with all 14 features, and got AUC score is 0.74. Although the AUC score is in acceptable range of 0.7 to 0.8, we tried to apply StandardScalar on the dataset to resize the distribution of variables so that mean of the observed value is 0 and the standard deviation is 1. We received slight improvement in the score to 0.83 (83%). Then, we tried to remove irrelevant features like capital-gain, capital-loss and fnlwgt from the dataset and ran Logistic regression with and without StandardScalar. We received the score as .8074 (80.74%) and .8108 (81.08%) respectively.

We also noticed the F1 score for the minority class (earning income > 50K) gradually increases when we ran logistic regression after applying StandardScalar and removing irrelevant features from the dataset.

**Decision Tree Classifier:**

We followed similar approach for decision tree classifier as we did for logistic regression, and received the accuracy as .8214 (82.14%) when we consider all the features from the dataset. This is slightly high compared to what we received for logistic regression. However, when we tried to run decision tree classifier on the dataset standardized using StandardScalar, the score didn’t improve much. We also noticed that score got decreased when we apply decision tree classifier algorithm on the dataset after removing irrelevant features.

We noticed the F1 score for the minority class (earning income > 50K) gradually decreases when we ran decision tree classifier model after applying StandardScalar and removing irrelevant features from the dataset. Initially, the F1 score for minority class is 0.62 and decreased to 0.53 during the subsequent modeling.

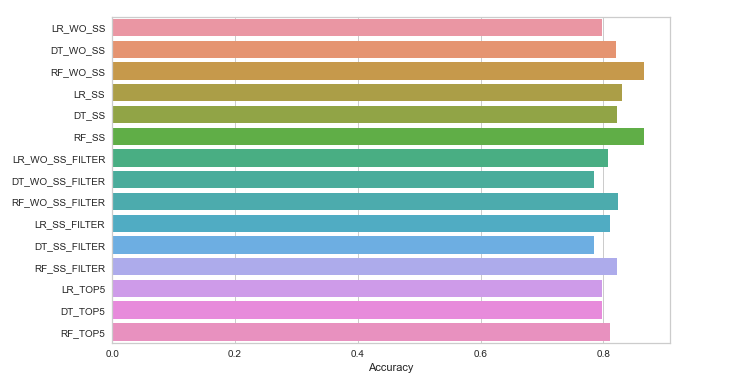
**Random Forest Classifier:**

Finally, we applied random forest classifier on the dataset with all 14 features included. We received the score as 0.8655 (86.55%) which is high compares to the score received for logistic regression and decision tree classifier. The same accuracy has been obtained after applying StandardScalar. Upon apply modeling on the dataset after removing irrelevant features, the score for random forest model has been reduced to 82.32%.

The F1 score for the minority class (earning income > 50K) gradually decreases when we ran decision tree classifier model after applying StandardScalar and removing irrelevant features from the dataset. Initially, the F1 score for minority class is 0.69 and decreased to 0.53 during the subsequent modeling.

Among all 3 models and multiple iterations, we noticed the AUC score is high for the random forest which is 0.91 when we run on the dataset with all 14 features. However, the score got reduced to 0.87 when we run on the dataset after removing unwanted features.

**Scores Plot**



*Figure: 4.5 Accuracy for different models applied on income dataset*

**Top 5 Features:**

Using SelectKBest, we tried to find the 5 best features after removing unwanted features from the dataset, and following are the best features in the dataset which shows higher impact to the target variable “income” compared to other features present in the dataset.

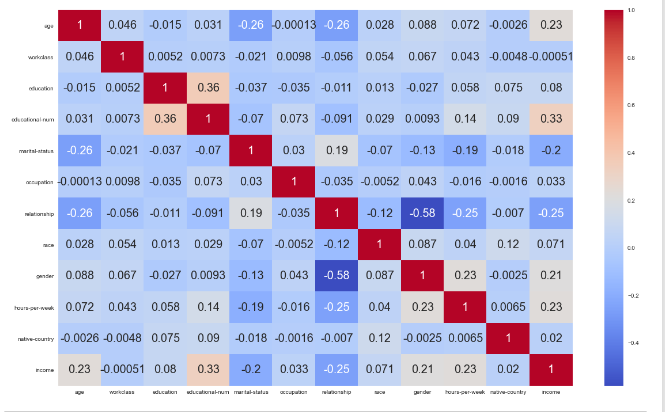
* Age
* Education-num
* Marital-Status
* Relationship
* Hours-per-week

Upon running logistic regression, decision tree classifier and random forest classifier model on the dataset with top 5 best features, the score turned out as 79.99%, 79.5% and 80.86% respectively. The corresponding AUC scores are 0.81, 0.77 and 0.85 respectively.

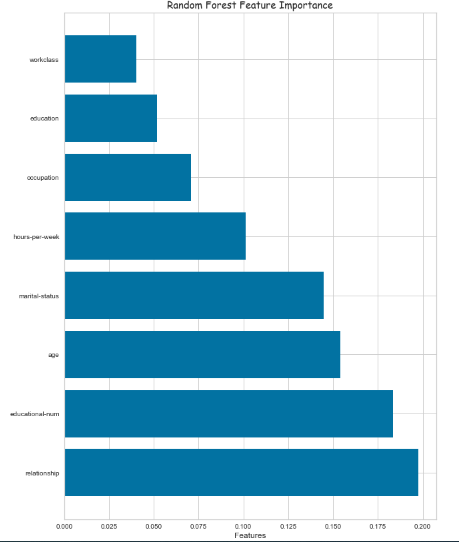
### **Begin to formulate conclusion/recommendations**

* A Logistic Regression Model, Decision Tree Classifier and Random Forest Classifier were used on the dataset to determine which features are most related to income of the household, where income is the target variable.
* Calculated the best performing features contributing to our target variable “income” of the household by following methods.

|  |  |
| --- | --- |
| **Test** | **Features** |
| Pearson's correlation matrix - Feature correlation to target variable "income" | **Numerical Variables:**  Age Education-num Hours-per-week  **Categorical Variables:** Gender Education |
| Chi-Squared (X2) Test - 5 Best features correlated to "Income" | Relationship education-num age marital-status hours-per-week |
| Using Feature Importance of Random Forest Classifier | Relationship education-num age marital-status hours-per-week |



*Figure 4.6: Pearson’s correlation matrix*



*Figure 4.7: Feature importance method of Random Forest*

**Findings & Recommendations:**

Out of three model, Random Forest Classifier is the best model to predict the income of the household as the score is higher compared to Logistic Regression and Decision Tree Classifier when we try to run the model for different scenarios.

Among various methods used to find 5 best features in the dataset, all the methods (Pearson’s correlation, chi-squared and feature importance of RF) provided age, education-num and hours-per-week as best features which have high impact on the target variable “income” compared to other features in the dataset. Pearson’s correlation gave Gender and Education as next best features, while chi-squared and feature importance of RF gave relationship and marital-status as best features having high impact.

From figure 4.7, following are the features having high impact of the target variable “income”

1. Relationship: Relationship of the person to the family
2. Education-num: Number of years the person had education
3. Age: Age of the person
4. Marital-status: Marital status of the person
5. Hours-per-week: Hour the person worked for a week

Following are some of the recommendations to earn high income.

1. You should work more hours per week to earn more income (Example: one who works 80 hours earn more compared to the one who works only 40 hours)
2. You should study for a greater number of years (number of education years) to earn high income.
3. The model also predicted that married persons earn more compared to unmarried persons. This may be due to the fact that persons with more experience (in term of years) is earning more compared to freshers. Mostly, the persons with more experience are married.
4. Model predicts that older people earn more compared to younger people. This implies that older people would have more experience who in turn earn more. So, you should get old to earn more income.

### **Reference:**

Abbott, D. (2014). *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*. John Wiley & Sons, Inc.

<https://www.census.gov/en.html>

<https://www.kaggle.com/datasets/uciml/adult-census-income>

<https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression>

<https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/binary-logistic-regression/before-you-start/data-considerations/>