CSC 8515 – Machine Learning
Final Report
Title: Classification of income using Adult Census Dataset
Professor: Dr. Ben Mitchell
Completed by: Nirajan Koirala

Table of Contents

Abstract

This project reports the analysis and exploration of the Adult Census Dataset and classifies the income levels of people using various demographic, social and economic features. Various Machine Learning techniques are employed to analyze the features present in the dataset. The dependency of income on each of these features is measured and this information is used to predict the income with the help of suitable algorithms. The correlation level of the features with income validated through these methods can help the authorities construct better plans for the future by helping to maintain healthy income levels.

1 Dataset

1.1 Source

The dataset named "Adult" is taken from the UCI Machine Learning Repository. The extraction of dataset was done by Barry Becker from the 1994 Census database. To make the dataset relevant to the working population, a set of reasonably clean records was extracted using the following conditions:
((AAGE > 16) && (AGI > 100) && (AFNLWGT>1) && (HRSWK>0)) where AAGE is the age, AGI is the adjusted gross income, AFNLWGT is sampling weight (demographic score based on residency and employment) and HRSWK represents hours-worked per week.

1.2 Overview

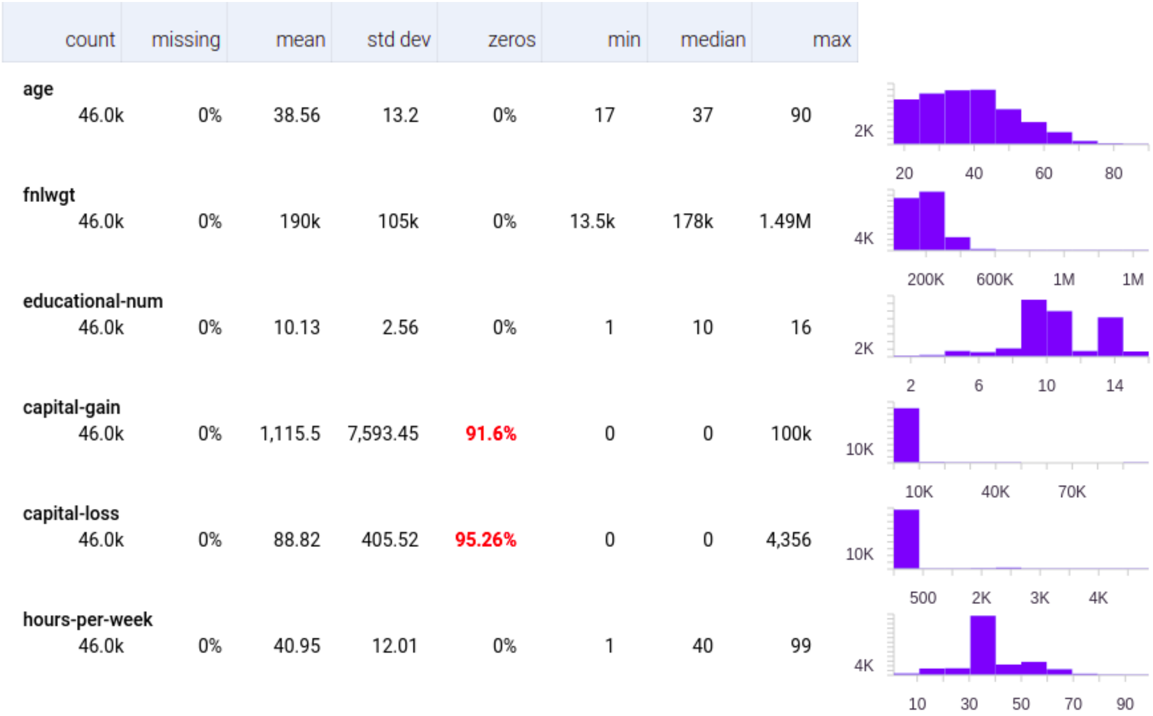
The dataset is made of 48,842 rows and 15 columns. Each row represents an individual's record and the columns represent the features of the dataset and the label, income. Income has a binomial characteristic having values as '<=50k' or '>50k'. The dataset contains 6 numerical features as age, fnlwgt, educational-num, capital-gain, capital-loss, hours-per-week and 8 categorical features as workclass, education, marital status, occupation, relationship, race, gender and native-country.

1.3 Features

1.3.1 Numerical features

The fnlwgt feature is the sampling weight which is a demographic score based on residency and employment. Individuals with similar demographic characteristics are expected to have similar fnlwgt values. Educational-num is the number of years of education obtained by an individual. Capital-gain and capital-loss represent the profit and losses respectively from investment sources apart from wages/salaries. Hours-per-week is the number of hours worked per week by an individual.

All the numerical features are normally distributed. Age and fnlwgt features are positively-skewed and educational-num is negatively-skewed. Hours-per-week has most of its values (roughly 80%) between 35-40. Age, hours-per-week, and fnlwgt have outliers in less than 5% of the instances. A high-kurtosis is found for capital-gain and capital-loss which indicates there is a high number of outliers for these features. More than 90% of the values are 0 in capital-gain and capital-loss.



	count	missing	mean	std dev	zeros	min	median	max
age	46.0k	0%	38.56	13.2	0%	17	37	90
fnlwgt	46.0k	0%	190k	105k	0%	13.5k	178k	1.49M
educational-num	46.0k	0%	10.13	2.56	0%	1	10	16
capital-gain	46.0k	0%	1,115.5	7,593.45	**91.6%**	0	0	100k
capital-loss	46.0k	0%	88.82	405.52	**95.26%**	0	0	4,356
hours-per-week	46.0k	0%	40.95	12.01	0%	1	40	99

Table 1 Distribution of numerical features

1.3.2 Categorical Features

Workclass feature describes the type of employer such as private, public, etc. Education describes the highest level of education attained which consists of high-school, bachelors, masters, etc. Marital-status of an individual consists of married-civ-spouse, divorced, etc. Occupation describes the employment type of a person which contains sales, craft-repair, tech-support, etc. Relationship is the type of relationship a person has in a family which consists of husband, wife, other-relative, etc. Description of race, gender, and native-country and other features are provided below.

Work-class :

- Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

Education :

- Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

Martial Status :

- Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

Occupation :

- Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

Relationship (individual's relation in a family) **:**

- Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

Race **:**

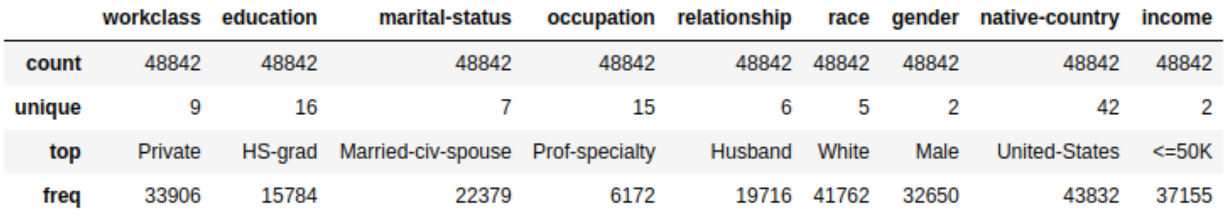
- White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

Gender **:**

- Female, Male.

Native-country **:**

- United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.



	workclass	education	marital-status	occupation	relationship	race	gender	native-country	income
count	48842	48842	48842	48842	48842	48842	48842	48842	48842
unique	9	16	7	15	6	5	2	42	2
top	Private	HS-grad	Married-civ-spouse	Prof-specialty	Husband	White	Male	United-States	<=50K
freq	33906	15784	22379	6172	19716	41762	32650	43832	37155

Table 2 Distribution of categorical-features

1.4 Label

The target column is the income which has values as ‘<=50k’ and ‘>50k’. The labels are not balanced. 76.07% values are <=50k and 23.93% values are >50k. The accuracy baseline for the models is set upon 76% based on these values.

1.5 Data Cleaning

Various data cleaning methods are applied to aid the visualization and label prediction process. There are no null values present in the dataset. 52 of the rows are found to be duplicates. All the duplicate rows are dropped to make the dataset contain unique row values. While checking for other inconsistencies, some rows contained relationship as husband and gender as female. Since we have only two types of genders present in the dataset, the rows with inconsistent relationship and gender values are dropped.

For the categorical features, three features are found with unknown values labeled as ‘?’. Workclass, native-country, and occupation contains 5.7%, 1.7% and 5.8% unknown values respectively. For workclass, unknown values are imputed with the median group, which is Private and it is present in almost 70% of the rows. For native-country as well the unknown values are imputed with the median value United-States since it was present in almost 91% of the rows. The unknown values in the occupation are dropped because the categories of occupation are highly distributed and the median category doesn’t represent a significant portion of the rows. For some of the categorical features, the existing groups are replaced into meaningful groups to aid the visualization and exploration process. Workclass is categorized into Private, Public, Self-Employed and Unemployed. Education is categorized into High-School dropout, High-School, Associate, Bachelors, Masters, Professional, and Doctorate. Marital-status is categorized into Never-married, Married, Widowed, and Separated. The categories in occupation and relationship are very distributed and its categories are not changed since no meaningful relationship between them could be found. Almost 86% of the rows have race feature as white. The data for other races is not enough to draw any valid conclusions for each of the individual races. Race is thus categorized into White and Non-White groups which might also help the classifiers in generalization. Similarly, native country is categorized into United-States and Non-United States as about 91% of its categories contain United-States. 45,981 rows are present in the dataset after the cleaning phase.

2 Exploratory Data Analysis and Hypothesis testing

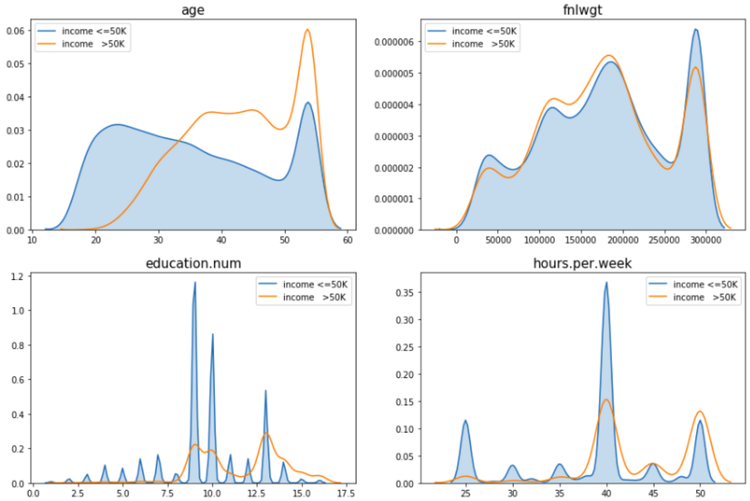
Each of the features except capital-gain, capital loss and fnlwgt are visualized with respect to income to see any potential correlation using various visualization techniques. Capital-gain and capital-loss have more than 90% of their values as 0 and fnlwgt has a significantly high range. All of their values are highly skewed and visualization of these features couldn't obtain any meaningful trends.

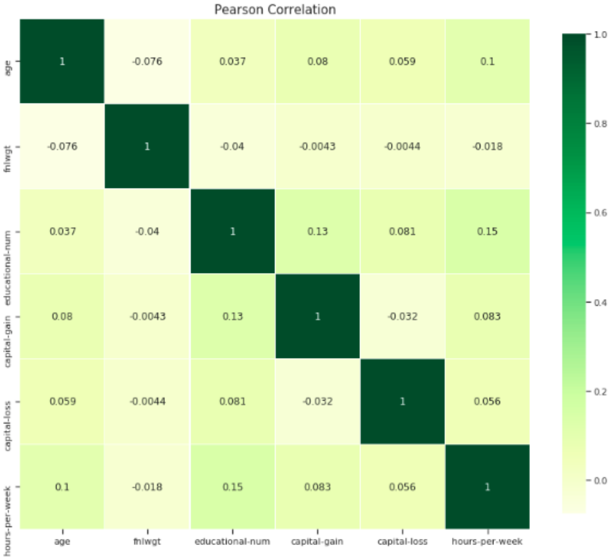
For the numerical features, Two Sampled T-test, also known as Independent t-test is used to perform hypothesis testing. This test is used as one categorical variable (income) with two groups and one continuously distributed numerical variable (feature) is used for testing. This experiment uses t-test values and p-values to reject/accept the null hypothesis. A statistical significance level of 0.05 is used. For the categorical features, Pearson Chi-square test is used to perform hypothesis-testing. This test is used as both the variables to be tested are categorical and they consist of two or more categorical, independent groups. The categorical test uses statistic and critical values to reject/accept the null hypothesis. The null and alternate hypothesis is defined as follows:

1) For the numerical features:
H_0: The two populations being tested are independent and they have no statistically significant different mean values.
H_a: The two populations being tested are dependent and there is a statistically significant difference present their mean values.

2) For the categorical features:
H_0: The variables being tested are independent and a non-significant result is found.
H_a: The variables being tested are dependent and a significant result is found.

2.1 Analysis of numerical features



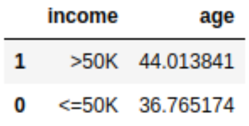


	age	fnlwgt	educational-num	capital-gain	capital-loss	hours-per-week
age	1	-0.076	0.037	0.08	0.059	0.1
fnlwgt	-0.076	1	-0.04	-0.0043	-0.0044	-0.018
educational-num	0.037	-0.04	1	0.13	0.081	0.15
capital-gain	0.08	-0.0043	0.13	1	-0.032	0.083
capital-loss	0.059	-0.0044	0.081	-0.032	1	0.056
hours-per-week	0.1	-0.018	0.15	0.083	0.056	1

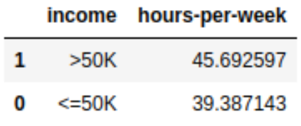
Figure 1 Distribution of Numerical Features

Figure 2 Heatmap-plot

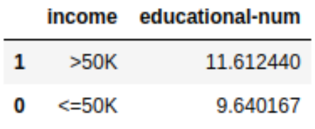
Using the heatmap we can observe that the higher values in the matrix correspond to the features in that row and column to be having some correlation with each other. For instance, the value 0.15 corresponds to educational-num and hours-per-week and indicates that these two features have a higher correlation than any of the other two features present in this table. None of the numerical features are found to have a strong correlation though. Mean values of age, hours-per-week, and educational-num of people who earn >50k are found to be greater than the mean values of the same group of people earning <=50k.



	income	age
1	>50K	44.013841
0	<=50K	36.765174



	income	hours-per-week
1	>50K	45.692597
0	<=50K	39.387143



	income	educational-num
1	>50K	11.612440
0	<=50K	9.640167

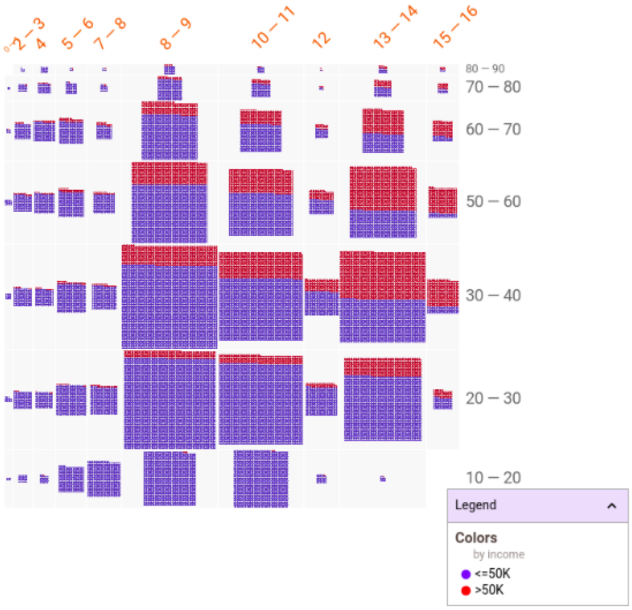
Hypothesis testing was carried out for these numerical variables. The testing supported our initial assumption of them being correlated since the p-value for all of them was found to be less than the significance level. Capital-gain was also found to be correlating with income. For capital-loss, the testing couldn't be carried out because nan p-values were produced. All the features correlating with income should help the classification models to predict income with high accuracy.

2.2 Analysis of categorical features

Within the workclass feature, Self-Employed people are found to have the highest proportion of people with >50k income. Similarly, married people in the marital-status, husband in the relationship, united-states in the native-country, exec-managerial in the occupation, white in race, male in gender are found to have a higher proportion of people with >50k income. These groups within the categories are assumed to have a potential correlation with higher income. To check the assumptions about the categorical feature's correlation with income, hypothesis testing was carried out for each of the variables. Using the chi-square test workclass, education, marital-status, occupation, relationship, and gender were found to have some dependency on income. Race and native-country were having no dependency on income.

2.3 Other Analysis Techniques

Analysis of most relevant features that were found to be correlating with income was carried out. To find the correlation of two numerical/categorical features with income, a visualization tool called Facets is used. Numerical features are transformed into several categorical buckets based on the range of values. Having this type of visualization helped dig more information about the income trends using both numerical and categorical features.



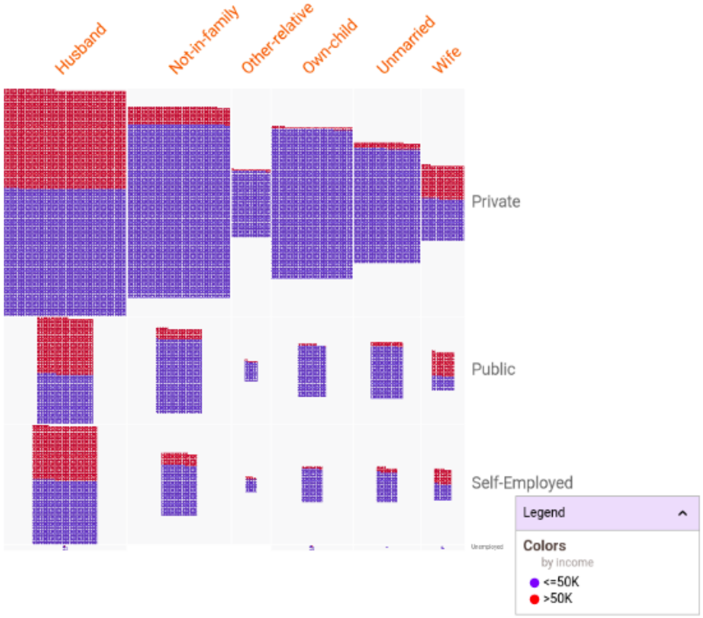
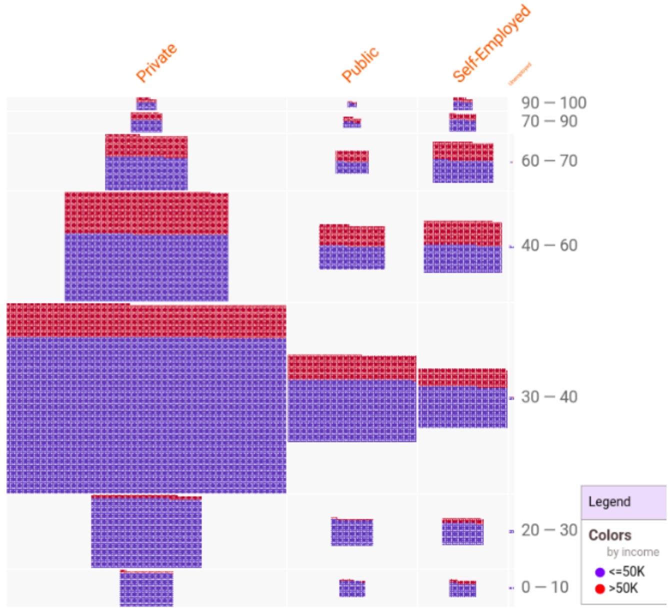


Figure 3 Educational-num and age

Figure 4 Relationship and workclass



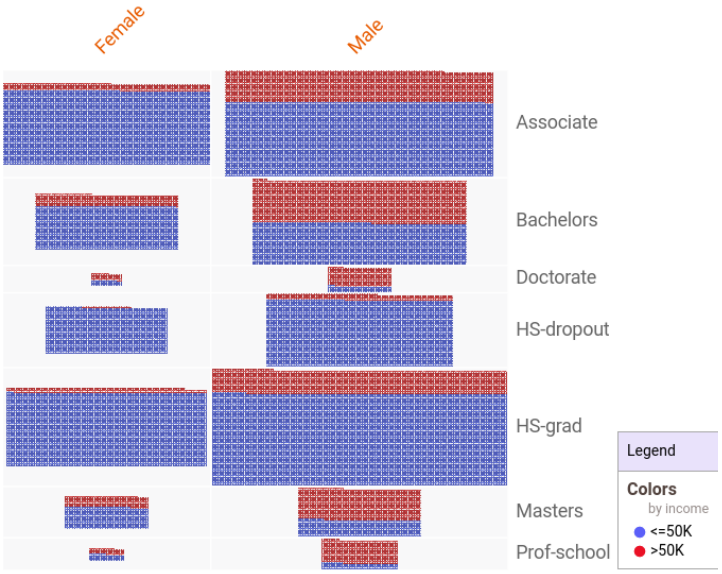


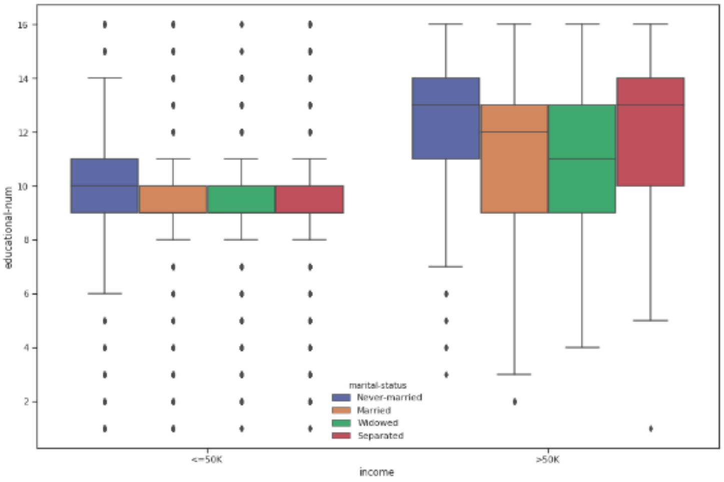
Figure 5 Workclass and Age

Figure 6 Gender and Education

Earlier during hypothesis testing, we observed that the number of years in school directly correlates with income. However, figure 3 shows that even though one may have higher educational-num values, his/her chances of having >50k seems to increase only after crossing about 30 years of age. Also, even though people may have mid-level educational-num values, as their age increase, the chances of them having >50k income increases. Prof-school having a lower educational-num than Doctorate has a higher proportion of people earning >50k. Most of the higher income-level groups were found to be located in buckets which are both higher in age and hours-per-week. A cross feature of these features is added in the feature columns to help the classifiers which generalize linearly.

Figure 3 further shows that only self-employed people are having no decline in the higher-income after 50-60 year buckets. This indicates that self-employed older people have better chances of earning >50k than their counterparts. One of the explanations for this kind of trend could be that self-employed people have steady income sources from their businesses and investments even after retirement. It was also observed that significantly more people in the private and self-employed category are working more hours than the ones in public. Figure 6 shows females have lower incomes in buckets where they have similar education as males. The median difference in ages of males and females gets higher in the higher income bracket of >50k as well where males have a significantly higher proportion. In figure 4, there seemed to be a decline in the number of people working in the self-employed group for all relationships except husbands. This suggests that husbands are more likely to be self-employed. Also, even though we see a disparity in the income between males and females in figure 6, one can observe that both husband and wife are earning higher than their counterparts in the relationship category and both have similar proportions in the >50k income groups in figure 4. This suggests married people have higher chances of earning >50k. Married marital-status seems likely to be a strong predictor for income after observing these trends.

Multivariate analysis techniques helped to observe some interesting trends.



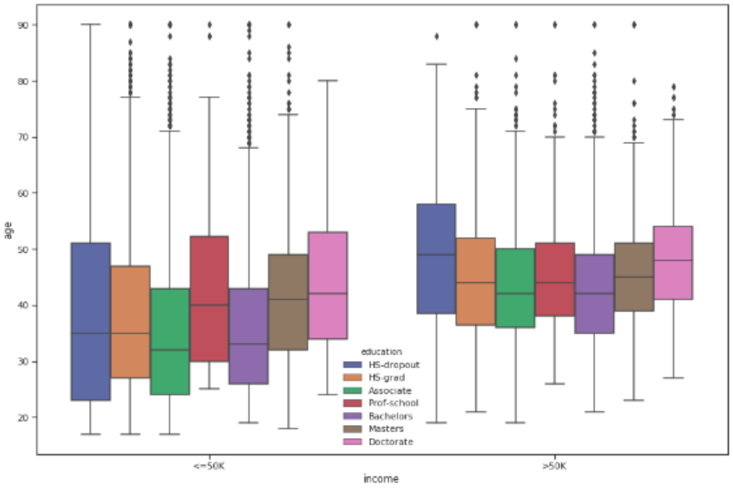


Figure 7 Education-num with marital-status

Figure 8 Age and Education

People who are never-married or separated and making >50k income have a significantly higher median number of educational-num than other categories. The median age of people who dropped from high school and have income >50k is relatively higher than other groups of education feature.

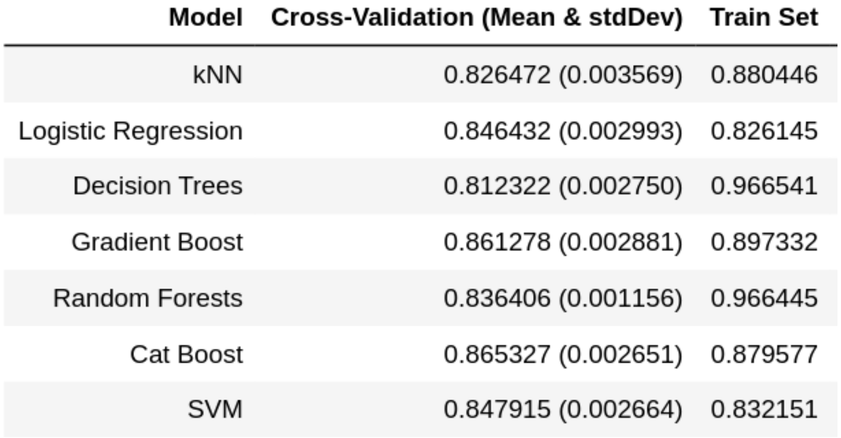
3 Methods used

3.1 Handling the features

After the visualization phase, age, education, and hours-per-week were found to be correlating with many of the other features in the higher income level groups. Two new features, education_age (education-num*age) and age_hours (age*hours-per-week) are added to the features columns. To select the most useful features, fnlwgt column is dropped based on very high p-values after the hypothesis testing and low Pearson correlation index values. Other features like race and native-country which were found to be independent with income, were dropped as well to help the classifiers generalize. Categorical features were transformed into sparse binary arrays using pandas get dummies. Since the numerical features had high range values, standard scaling was performed on them and they were normalized between values -1 and 1.

3.2 Validation and training

To deal with the imbalance amount of labels, the resampling method was used. Over-sampling of the minority label, >50k is done using SMOTE (Synthetic Minority Oversampling Technique). This method decreased the accuracy levels of some classifiers but the F1-score on the positive labels is improved. Stratified-fold cross-validation with 3 splits is used to validate the models. Due to the imbalanced proportion of labels, this technique helped ensure that each fold contains roughly the same proportion of class labels. The dataset is divided into 80% train set and 20% test set and final models are built.



Model	Cross-Validation (Mean & stdDev)	Train Set
kNN	0.826472 (0.003569)	0.880446
Logistic Regression	0.846432 (0.002993)	0.826145
Decision Trees	0.812322 (0.002750)	0.966541
Gradient Boost	0.861278 (0.002881)	0.897332
Random Forests	0.836406 (0.001156)	0.966445
Cat Boost	0.865327 (0.002651)	0.879577
SVM	0.847915 (0.002664)	0.832151

Table 3 Accuracy scores on the cross-validation and train set

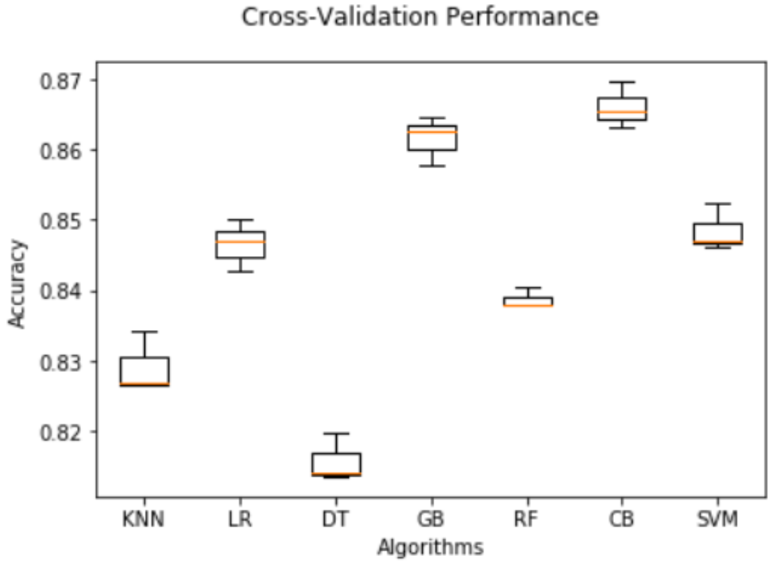


Figure 9 Box-plot for the cross-validation scores

3.3 Evaluation methods

Evaluation of the models is done based on confusion matrices, accuracy levels on cross-validation, train and test set, and F1-scores. F1-score is computed using the harmonic mean of precision and recall values.

$$F_1 = \left(\frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}\right) = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

F1-score is one of the most robust performance metric evaluation of these models since almost 76% of the rows have <=50k income labels. ROC curve is also an option while dealing with unbalanced datasets like this but since >50k classes are rare and we care more about the false positives than false negatives, F1-score is chosen over ROC. Tuning different hyper-parameters of the models is carried out to choose the parameters with the best F1-scores and accuracy levels. This process is time-consuming and to determine the best model, tuning all of the available hyper-parameters on the models was not possible. Top 3 models were chosen and major tunings were applied to these models to further improve their performance metrics.

Dimensionality reduction was carried out using PCA to increase the overall speed-performance of the models as the dataset contained 44 features after transforming the categorical columns into sparse arrays. 95% of the variance in the data was explained by the first 21 components. After applying PCA, the classifiers had average scores on the cross-validation however on the test set, most of them performed poorly. All the classifiers except KNN and logistic regression performed significantly poor on the test-set after PCA. KNN and logistic regression are linear classifiers and reducing the number of dimensions helped increase their performance. Since PCA transformed all of the features into different components, other models were not able to detect any correlation between the transformed components. Since applying PCA was not the best idea in building final models, it was not applied.

4 Models

4.1 Algorithms

4.1.1 K Nearest Neighbors

KNN is a linear classifier and it was expected to perform well by finding any linear correlations between the features. It over-fitted on the train set and performed average on the test struggling on the >50k labels. Different values of k were used on the test set to find the best performance but such type of methods is not applicable to real-world test data.

4.1.2 Logistic Regression

Logistic regression is one of the widely used binary classifiers. Since it uses a linear method for classification, it was able to capture most of the linear correlations between the features and income. It ‘s performance was average on both training and test sets.

4.1.3 Decision Trees

Decision trees are a non-parametric supervised learning method used for classification. It was expected to capture most of the correlations between the features but due to low number of higher-income labels, it couldn’t capture this correlation effectively. It highly overfitted the training set and its performance of test sets was average. Different algorithms like gini and entropy for building the trees were experimented with but other methods like pruning was not carried out.

4.1.4 Gradient Boosting

Gradient boosting works by using decision trees. These trees are called weak learners and different constraints are put like minimizing/maximizing the depth, nodes and splits in the trees to perform classification. It was able to perform well obtaining high results on both train and test sets.

4.1.5 Random Forest

Random Forest is an ensemble learning method for classification and operates by constructing a multitude of decision trees. It highly overfitted the train set and performed average on the test set.

4.1.6 Cat Boost

Cat Boost is based on gradient boost algorithm. It works well with datasets containing a large number of categorical columns. It performed the best among other classifiers resulting in very high train as well as test set scores. Also, it had the highest F1-score for >50k label on the test set.

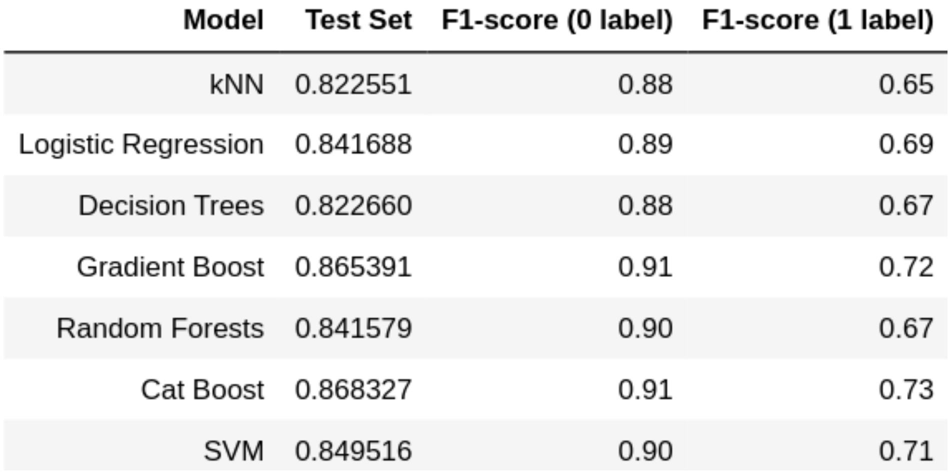
4.1.7 Support Vector Machines

SVMs were used for modeling without any fine-tuning of the parameters. Using rbf kernel it produced high train as well as test scores without any overfitting.

4.1.8 Artificial Neural Networks

ANN models with different amounts of layers were also experimented with and built. RELU was used as the activation function for hidden layers. Its performance was compared with other algorithms and the best accuracy score achieved was 84.17% but other metrics like cross-validation and F1-scores for the labels were not evaluated.

4.2 Comparison of models



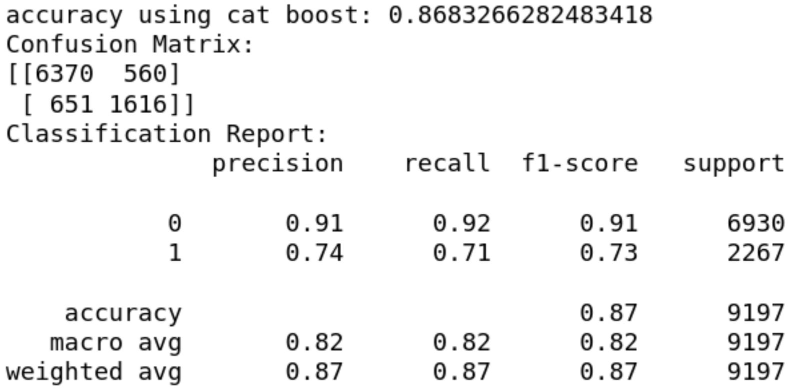
Model	Test Set	F1-score (0 label)	F1-score (1 label)
kNN	0.822551	0.88	0.65
Logistic Regression	0.841688	0.89	0.69
Decision Trees	0.822660	0.88	0.67
Gradient Boost	0.865391	0.91	0.72
Random Forests	0.841579	0.90	0.67
Cat Boost	0.868327	0.91	0.73
SVM	0.849516	0.90	0.71

Table 4 Accuracy scores on test set and F1-scores

All the models are performing higher that the set baseline accuracy. Every model is performing good on the 0 label (<=50k income) due to sufficient training samples. For gradient boost, a high test-set score was achieved through experimenting and changing different hyper-parameters like the learning rate and depth of the tree. However, it would not be possible to customize the parameters differently for train and test sets, especially in a real-world setting where one will have no information about the test set.

SVM model was able to get a very high test set score and good F1-scores. Cat Boost outperformed all other models for classifying income obtaining high cross-validation, train, test scores, and a high F1-score especially on the 1 label (>50k income). It also showed no signs of any overfitting.

5 Results



```
accuracy using cat boost: 0.8683266282483418
Confusion Matrix:
[[6370  560]
 [ 651 1616]]
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.92      0.91      6930
           1       0.74      0.71      0.73      2267

    accuracy                           0.87      9197
   macro avg       0.82      0.82      0.82      9197
weighted avg       0.87      0.87      0.87      9197
```

Table 5 Report for Cat Boost

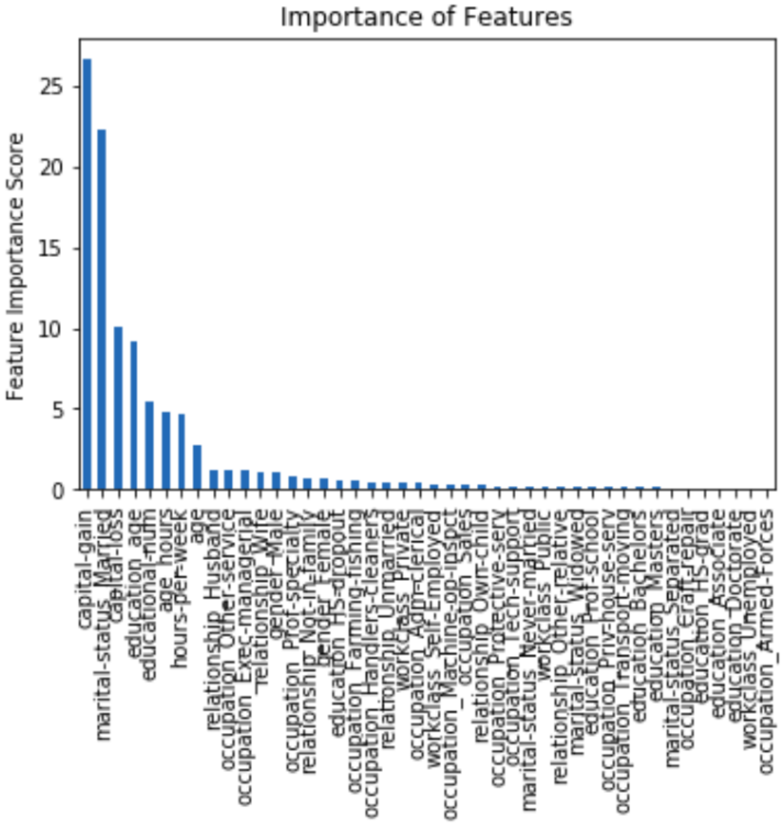
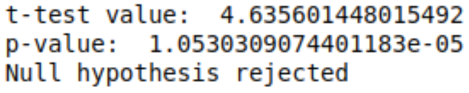


Figure 10 Best Features using Cat Boost

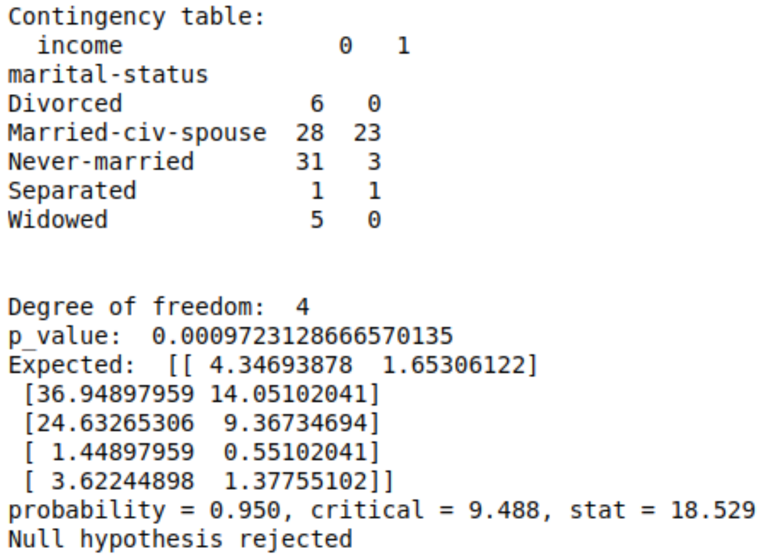
The most useful features for classifying income are determined using Cat Boost. The graph depicts capital-gain, martial-status_married, capital-loss and education-age as the best features for predicting income. Relationship_husband and relationship_wife are also higher on the importance scale which suggests that married marital-status is almost equal to if not greater than capital-gain on the feature importance score. Surprisingly many factors like education, age and occupation are not able to come higher on the importance scale.

Hypothesis testing results for Capital-gain using independent t-test:



```
t-test value:  4.635601448015492
p-value:  1.0530309074401183e-05
Null hypothesis rejected
```

Hypothesis testing results for marital-status using chi-square:



```
Contingency table:
  income              0   1
marital-status
Divorced            6   0
Married-civ-spouse  28  23
Never-married       31   3
Separated           1   1
Widowed             5   0


Degree of freedom:  4
p_value:  0.0009723128666570135
Expected:  [[ 4.34693878  1.65306122]
 [36.94897959 14.05102041]
 [24.63265306  9.36734694]
 [ 1.44897959  0.55102041]
 [ 3.62244898  1.37755102]]
probability = 0.950, critical = 9.488, stat = 18.529
Null hypothesis rejected
```

All of the features who were found to be independent of the income label during hypothesis testing have very low importance scores as well which validates the previous results.

6 Challenges Encountered

One of the challenges faced was finding meaningful trends during visualization. Since the data is collected from the 1994 census, it was difficult to find correlation and income trends that would be applicable in today's world. Visualizations were performed before transforming the categorical features into numerical values. This step helped because direct observations could be made for the categorical features and their groups. However, it proved to be difficult to plot categorical columns against numerical ones. Several techniques were tried for this task like bucketing the already present numerical columns into categorical buckets but this method was also only limited to plotting only 2 features against income. It was a lengthy process to observe the trends and come up with meaningful conclusions from the patterns. Hence, the visualization of multiple interactions between several features and the income trends was skipped out. Data is also suffering from selection bias. Almost 90% of people are white and from the United-States. The analysis and trends obtained for the income might be thus only applicable to these groups of people. The unbalanced amount of labels had to be mitigated as well using techniques like SMOTE before training the classifiers.

7 Conclusion

This project helped extrapolate many types of different patterns from the census data. The accuracy levels and F1-scores obtained from the best models are also very promising. Feature importance graph helped observe the most important features for having higher income levels. As the data suffered from imbalanced labels, boosting models performed well on the rare positive classes after oversampling. Further improvements in the models can be made by fine-tuning different parameters for each of the models built. More meaningful and modern trends in income could also be captured if similar visualization techniques and analysis could be applied on a relatively new census dataset. These results shows that income correlates more with many non-obvious factors like marital-status and age instead of obvious factors like occupation, workplace, and even education. Digging these type of correlations with income can help one make plans for a society which would help in the future to maintain healthy income levels of people.

8 References

[1] https://www.kaggle.com/ipbyrne/income-prediction-84-369-accuracy

[2] https://towardsdatascience.com/logistic-regression-classifier-on-census-income-data-e1dbef0b5738

[3] https://www.kaggle.com/jieyima/income-classification-model

[4] https://medium.com/all-things-ai/in-depth-parameter-tuning-for-gradient-boosting-3363992e9bae

[5] https://www.kaggle.com/kanav0183/catboost-and-other-class-algos-with-88-accuracy