CS777 – Term Project Report

**Student: Andrew Marin**

# Introduction:

# *Data set description:*

The dataset that I have selected is a public dataset from Kaggle in .csv format and it is an Adult Census Income dataset with ~ 32,000 records.

<https://www.kaggle.com/datasets/lovishbansal123/adult-census-income/data>

This dataset contains demographic and employment information from the U.S. Census Bureau, including age, work class, education level, marital status, occupation, relationship, race, gender, hours per week worked, and native country. It also includes an ‘income’ column, equivalent to a class label, indicating if the individual in that row’s income surpasses $50K/year. I believe this dataset is originally from 1994, as I found a similar dataset from UC Irvine with that date label. The dataset contains 15 columns including the class label, 9 of which are strings and 6 of which are integers. All the personal data in the dataset has been anonymized.

# *Research Question:*

My research question is: Can we accurately predict whether an individual's income exceeds $50,000 per year based on their demographic and employment characteristics? I believe this is important because studying the factors that influence income is crucial for several reasons: It allows us to draw socioeconomic insights so that we can understand demographic factors that contribute to higher or lower income and potential areas of disparity within our society. It can affect resource allocation because government and local leadership are able to identify groups of individuals that may need additional support services. It can also shape how we look at career guidance and education because if there are related factors associated with earning higher incomes, we want career counselors to be able to advise individuals on career paths / skills development that will lead to higher earning potential. I am hoping to learn which demographic and employment features contribute the most to predicting whether an individual earns above or below 50K dollars per year. I also want to see the relative importance of the remaining factors (perhaps some are not important at all). I hope to see strong overall accuracy and effectiveness of a classification model I build in predicting income.

# Methodology:

# *Data Cleanup & Preparation:*

I load the data into a spark data frame and trimmed any extra white space in the string columns. I then checked for missing values, which there are none in the original dataset. However, the original dataset contained ‘?’ within the data to represent missing values, so I substitute ‘?’ with ‘Unknown’.

I then look through each of the categorical variables in my dataset and limit the number of unique entries within each category to the top 10, including ‘Other’, which is used to group the remaining distinct values. I did this because using decision trees requires less than a certain number of possible nodes for each category. I encode the categorical features using indexers and create a pipeline to fit and transform the data. I finally convert the binary class label ‘Income’ to be 1 (positive) if the string value was ‘>50K’ and 0 (negative) if the string value was ‘<50K’.

# *Feature Selection & Numeric Feature Scaling:*

In this stage, I use a Random Forest model to aid in my feature selection. I assemble my features into a vector and train the Random Forest model. I then print out the feature importances and subset my data set to only include the top 10 features. Afterwards, I scale the numeric features using the MinMax scaler and drop any of the old feature columns so that prior to running the ML models, my dataset is only the feature vector of my numerical and categorical features, as well as my binary class label column.

I scaled my numeric features after deciding on which features to include because Random Forest does not use distance metrics, so I did not need to scale prior to deciding, and I wanted to the keep the individual level of detail for each column before collapsing them into a vector. This way, I can determine which features contributed the most to predicting the class label.

*Machine Learning Models:*

For this project, I used a Logistic Regression model and Decision Tree model to act as classifiers on my dataset. My objective was to build a model that can predict with high accuracy if an individual’s income exceeds $50,000 (binary outcome). I split my data into training and testing subsets using an 80/20 training-testing split and I prepared parameter grids to test different values for regularization and max-depth in Logistic Regression and Decision Trees, respectively. I then set up my cross validation for both models so that the training data will be subset into 5 folds. Finally, I fit the cross-validated models onto the training data.

*Evaluating the Models:*

I evaluated the machine learning models performance on the test subset of data using accuracy, precision, recall, F1 score, and the confusion matrices that result from each ML model. I use the ‘MulticlassClassificationEvaluator’ package to generate those metrics along with the Multiclass Metrics package to prepare the predictions and labels for viewing in the confusion matrix. I compared the results of the two models against each other and used those findings to make any adjustments to the feature selection or model parameters.

# Results:

For my results, I tested my models a few times with slight differences between them. The code included in this assignment is from my final iteration. The first major factor I changed was the number of features I included in my models. I started with the top 10 based on Random Forest importance and reduced that number to the top 5 to determine if there was a significant effect on results.

***10 Features:***

Top Features from Random Forest in order:

1. Feature: marital\_status\_index, Importance: 0.2651945030523563
2. Feature: relationship\_index, Importance: 0.19880143578932652
3. Feature: capital\_gain, Importance: 0.18112189271794532
4. Feature: education\_num, Importance: 0.1535848390798236
5. Feature: occupation\_index, Importance: 0.07256028669861023
6. Feature: age, Importance: 0.04429071725741498
7. Feature: hours\_per\_week, Importance: 0.027620319802384812
8. Feature: capital\_loss, Importance: 0.024507358275500553
9. Feature: education\_index, Importance: 0.01887003877203624
10. Feature: sex\_index, Importance: 0.011639916229644762

**Logistic Regression Metrics:**

Accuracy: 0.8371626831148805

Precision: 0.8285629835186209

Recall: 0.8371626831148805

F1 Score: 0.8237138299105652

**Decision Tree Metrics:**

Accuracy: 0.8507324595219737

Precision: 0.8451702420395162

Recall: 0.8507324595219738

F1 Score: 0.8466729914101241

**Confusion Matrix for Logistic Regression:**

[[4675. 247.]

[ 809. 754.]]

**Confusion Matrix for Decision Tree:**

[[4549. 373.]

[ 595. 968.]]

***5 Features:***

Top 5 features from above

**Logistic Regression Metrics:**

Accuracy: 0.8331534309946029

Precision: 0.8240900804265774

Recall: 0.8331534309946029

F1 Score: 0.8181305293776462

**Decision Tree Metrics:**

Accuracy: 0.8427139552814187

Precision: 0.8345456758461279

Recall: 0.8427139552814187

F1 Score: 0.8345008245991968

Confusion Matrix for Logistic Regression:

[[4681. 241.]

[ 841. 722.]]

Confusion Matrix for Decision Tree:

[[4607. 315.]

[ 705. 858.]]

***Additional Testing & Handling Class Imbalance:***

I opted to use 10 features because we see a decrease in the metrics when trying the top 5 features. I also tested additional changes to my options for hyperparameters in the parameter grid so that I can test different combinations of the regularization parameter and iterations for Logistic Regression and different combinations of max depth and max bins for the Decision Tree model during my cross-validation. Below are the results of the code that is included with this project:

**Logistic Regression Metrics:**

Accuracy: 0.8371626831148805

Precision: 0.8285629835186209

Recall: 0.8371626831148805

F1 Score: 0.8237138299105652

**Decision Tree Metrics:**

Accuracy: 0.8556669236700077

Precision: 0.8510424814187109

Recall: 0.8556669236700077

F1 Score: 0.8524776054789427

**Confusion Matrix for Logistic Regression:**

[[4675. 247.]

[ 809. 754.]]

**Confusion Matrix for Decision Tree:**

[[4546. 376.]

[ 560. 1003.]]

# Discussion:

# ***Model Performance Comparison:***

Based on the output of both models in the final run, we can see that although the overall accuracy of both models was very close, the Decision Tree classifier had slightly higher metrics across the board when compared to the logistic regression model. We can see that the decision tree model had more false positives (376) than the logistic regression model (247) but it also had less false negatives (560) than the logistic regression model (809). The decision tree model was ultimately able to predict more true positives, so if we prioritize correctly classifying true positives, the decision tree model is more efficient. In this project, I wanted to create a model that was successful in identifying individuals who made over $50K a year, and I believe that the decision tree model does a good job of doing that. There is class imbalance in my original dataset which is exemplified by the large numbers of false negatives in each model’s results. If I were to go forward trying to improve one model, I would pick the decision tree model.

# ***Feature Importance:***

The top 10 features based on my Random Forest classifier were as follows:  
Feature: marital\_status\_index

Feature: relationship\_index

Feature: capital\_gain

Feature: education\_num

Feature: occupation\_index

Feature: age

Feature: hours\_per\_week

Feature: capital\_loss

Feature: education\_index

Feature: sex\_index

These top features coincide with what I expected because they seem reflective of an individual’s socioeconomic and personal background. If were to use common rationale to examine the above features, it would make sense that individuals with a higher number of education years, greater number of hours per week worked, and higher capital gains would be more likely to make more than $50K a year. This study is also from 1994, which leads me to believe that individuals with a sex of male were also making more money than females on average. Given that the categorical features did not have a finite set of answers to choose from and I had to limit the values within each category to the top 10 by count, we may be missing out on an extra level of detail within the values of those categories that were grouped together. The results from looking at the top 10 features vs. the top 5 features also pointed to better results when including all top 10 features.

# ***Limitations:***

This dataset is outdated (1994) any may have potential biases based on the time or way the data was collected. We also use logistic regression modeling which does assume that there are linear relationships present among the data. The given data does have evidence of class imbalance, and while I was testing adding weight columns to the training and testing data based on frequency, I visualized slightly worse performance (79% accuracy). This could be due to overcompensation for the class imbalance where we were overfitting on the minority class.

# Conclusion:

Overall, the models that I created were successful in accurately predicting whether an individual makes more or less than $50K a year. The Decision Tree classifier was better at predicting true instances of individuals making more than $50K a year, and the most important features based on the Random Forest model importance were an individual’s marital status, relationship status, capital gain, # of years educated, their occupation, age, and # of hours worked per week. By understanding the factors that influence a person’s salary, we can use these models to test new entries that will predict with a high level of accuracy if an individual makes more than $50K a year.

I think both models do a good job at balancing precision and recall based on the F1 score. The models accurately identify individuals who make over $50K a year while avoiding too many false positives. The models seem balanced which is important given that the class distribution in the original dataset is imbalanced. I would still choose the Decision Tree model going forward given that it had the highest true positive rate, and the focus of this project was to be able to identify individuals who make over $50K a year.

I believe that I could improve the models by incorporating additional feature examination to remove any redundant features before doing my feature selection with a random forest model. I also think that I need to re-examine how best to apply weights to my models. Although the results don’t show skewed results based on class imbalance, I may be able to increase the accuracy of the results even more by finding the right weights to apply. Other options include increasing the number of folds for my cross-validation or including more parameters to tune in my parameter grid when running cross-validation.