Adult

Dosbol Aliev
daliev@cougarnet.uh.edu

Adil Iqbal
asiqbal@cougarnet.uh.edu

Angelita Krepel
angelkrepel@gmail.com

Udochukwu Amaefule
Umaefule@cougarnet.uh.edu

Malik Taylor
Mtaylor@cougarnet.uh.edu

Dr. Cathy Poliak
cpoliak@central.uh.edu

ABSTRACT

The dataset our group selected is called ”Adult.” This is a dataset that collected data based on a census that provides information about adults to help us better understand what qualities could contribute to the outcome we are predicting. We got inspiration from this from looking at the UCI-Machine Learning Repository and because we are all interested in what factors of an adult’s life could affect salary.

SUBJECT

The size of the data is 48842 observations and 14 variables. The descriptions of the variables are: age: continuous, work-class: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked, fnlwgt: continuous, 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, education-num: continuous, marital-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: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried, race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black, sex: Female, Male, capital-gain: continuous, capital-loss: continuous, hours-per-week: continuous, native-country: country they were born, class: if person makes 50K or more. The variable that will be used as the response will be the class variable because we will be using that to predict whether or not the features will affect (positive/negative) if an adult makes 50k usd or more. To determine which variables are going to the predictors, we will be looking at different plots and diagrams to determine which ones will affect the response most. We will ask whether the working class the adults belong to affects how much they make, whether the country in which they were born has an effect, and so on.

EVALUATION

A successful outcome for this project is to use some machine learning models to get higher than 70 percent accuracy. To get such high accuracy, it will be used some machine learning tools to minimize training accuracy while maximizing the testing accuracy. Also, in this project will be used Hyper-turning parameters, increasing our chances of getting the highest precision to predict whether a patient has diabetes or not.

RELATED WORK

Since our response variable is binary and classifying, we will use logistic regression and probably decision trees. Using logistic regression is good because it is easier to implement machine learning methods, especially when it comes to testing and training sets. It also uses computational power and is suitable for linearly separable datasets, such as ours, where you can draw a straight line through the classes. The advantage of using decision trees is that they are simple to understand and interpret and require little data preparation. The cost of using the tree is logarithmic in the number of data points used to train the tree. It can handle both numerical and categorical data. We will also use cross-validation to see which of the two models is better for our dataset. Adil, Dosbol, and Angelita will be working on the data preparation, ensuring there are no nulls/missing values and that if there need to be dummy variables, those are created, as well as the cross-validation and comparisons of the models. Malik, Udochukwu, and Angelita will be working on their models and implementation. We will all focus on the data questions we have and ensure our models best suit the dataset.

CONCLUSION

We will use the Adult dataset to determine whether we can predict, with at least 70 percent accuracy, whether someone will make greater than 50,000 usd or not. This is a prediction problem and a classification problem. Though we have not settled on which models we will use, we will likely use logistic regression and decision trees. We will divide the workload among our team members as outlined in this document while also relying on each member to fill any gaps we have not yet foreseen. Along with our visualizations, we will also provide an interpretation of the results of our experiment.