Alex Nanez

Dr. Rostami

DATA 3421

1 May 2023

Final Project Report

Introduction

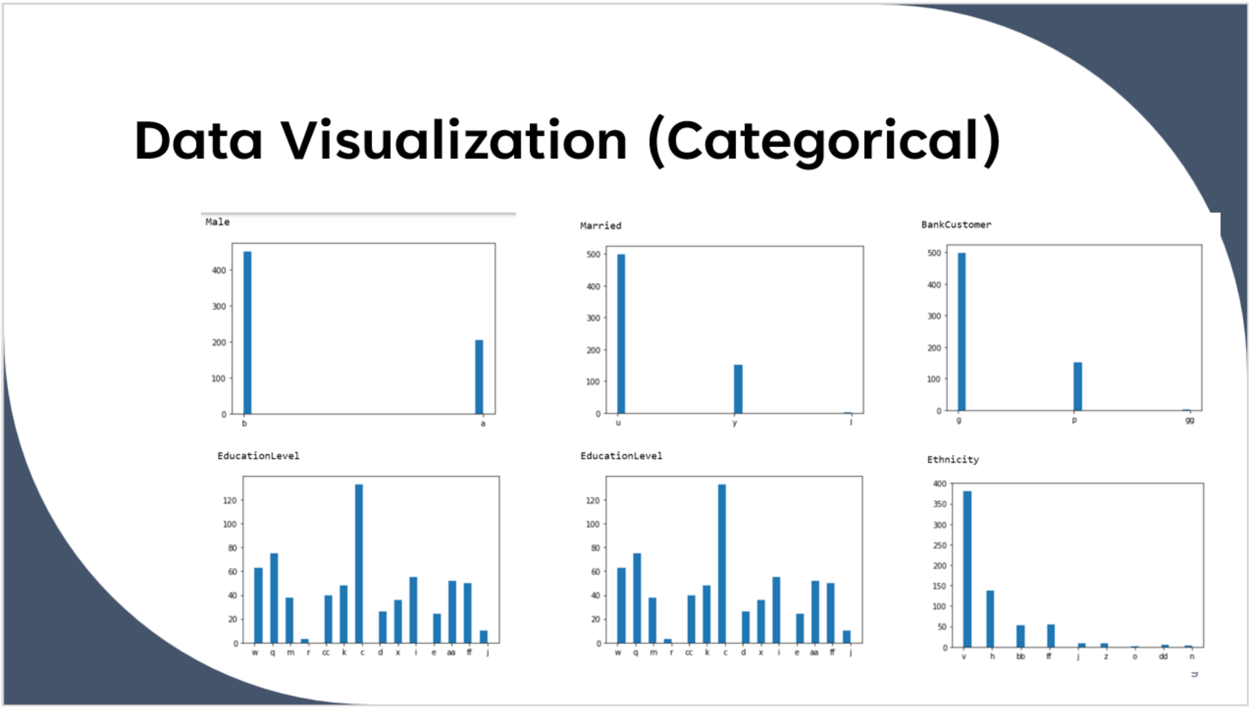
One important part of the financial industry is managing credit card applications and deciding whether a credit card will be issued to an applicant. Most banks have a system that does this automatically since it would be arduous and repetitive to manually review millions of credit card applications. My focus is on constructing machine learning models to predict this decision, and I used the Credit Approval Data Set from the UCI Machine Learning Repository for my project. There was a total of 690 instances and 16 attributes, including the class attribute.

Data Preparation and Preprocessing

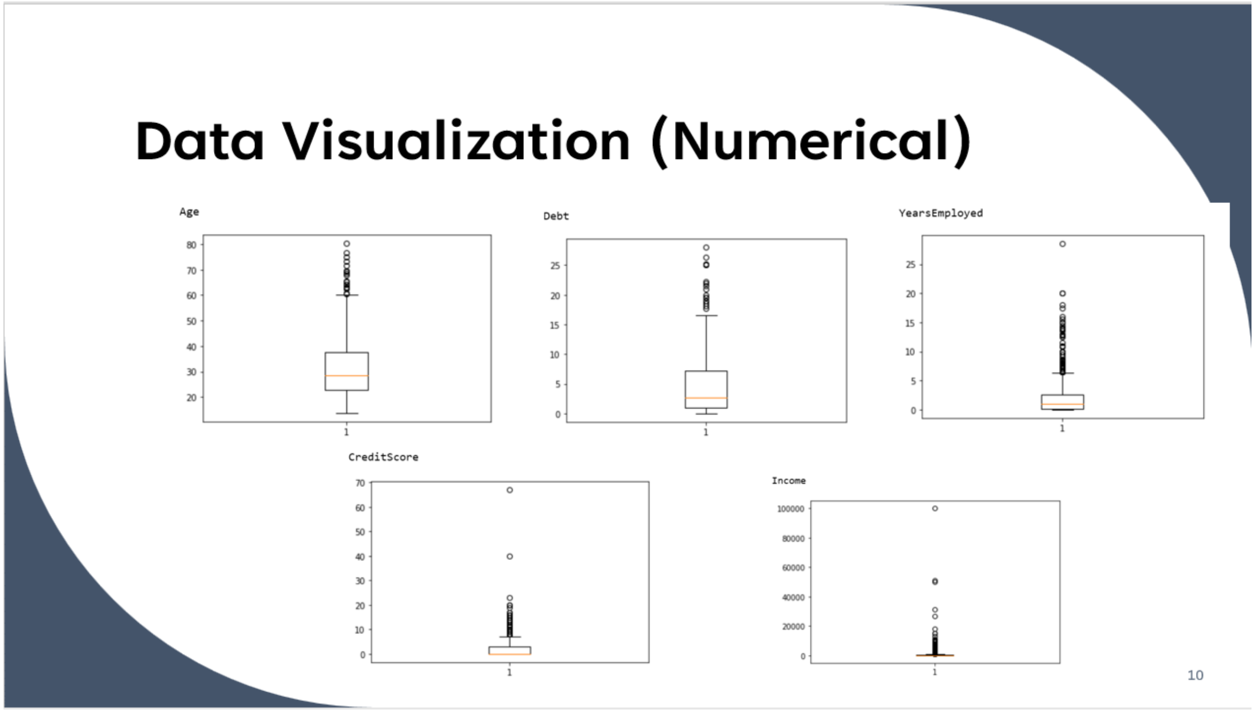
One of the biggest challenges of this data set is that it was encrypted to protect the confidentiality of the data. Thus, one step of data preparation I decided to take was give all 16 features working names to make it easier to work on the data. There were 37 rows (5%) with missing values that needed to be dealt with, and they were represented by question marks in the original data set. After converting them to None values in pandas, I decided to perform data imputation to replace the missing values based on their classification. I used the median as what would be replaced for the numerical feature with missing values and the mode for the categorical features. Finally, before implementing my machine learning models, I performed feature scaling using scikit-learn’s MinMaxScaler method as well as label encoding using the OrdinalEncoder method.

Exploratory Data Analysis

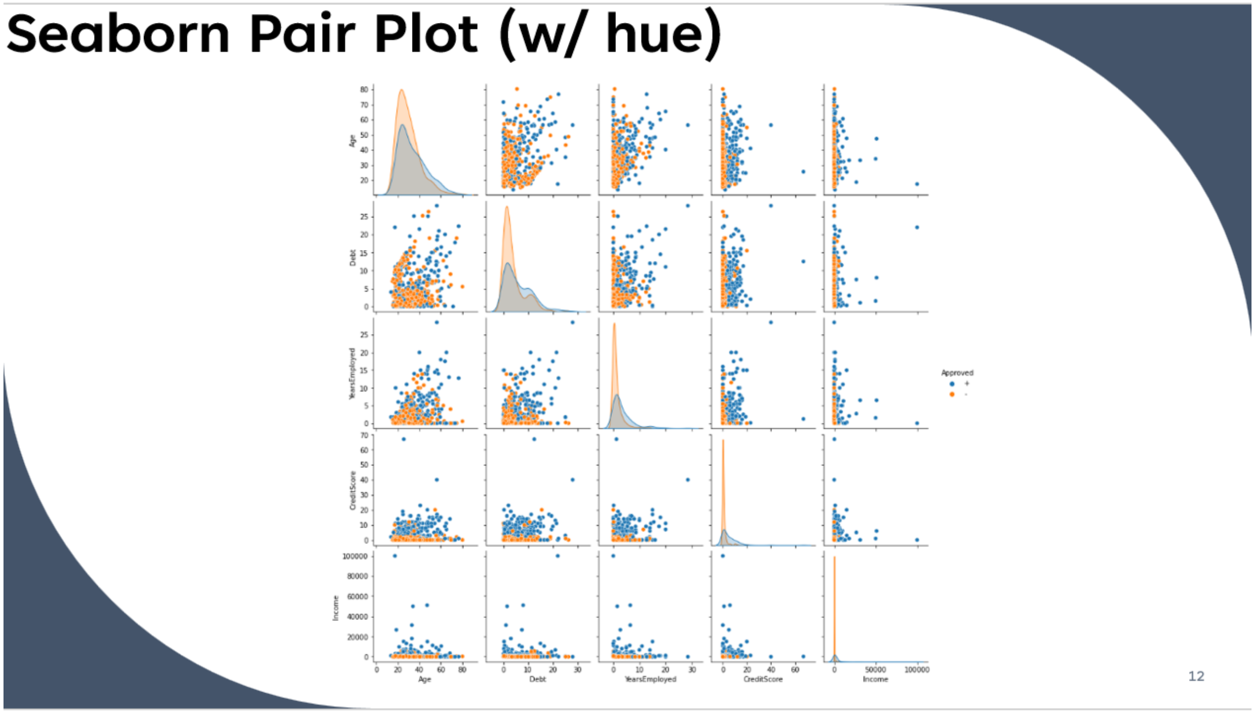
When the data was cleaned, I looked at the summary statistics of the data and performed data visualization. For the categorical variables, I used histograms to better look at the distribution, and I could see which encrypted values were most frequent. I used boxplots for the numerical variables and found that all of them were positively skewed, so scaling the features was essential before performing the models. I also made a pair plot using Seaborn while including whether a data point was approved or not.



Data Visualization (Categorical)



Data Visualization (Numerical)



Seaborn Pair Plot (w/ hue)

Model Selection and Training

For this project, I decided to employ 3 machine learning algorithms, which including logistic regression, decision tree, and K-Nearest Neighbors (KNN). Before training the data, I defined different variables for the feature and target variables. Then, I split my data into training and testing sets using a 70:30 ratio and started training my data using each model.

Model Evaluation and Results

For logistic regression, the test accuracy I obtained from the confusion matrix was 83%. For decision tree, my initial test accuracy was 75%, but after finding the optimized ‘max_depth’ hyperparameter (1 or 2), my final test accuracy was 82.6%. For KNN, I optimized the ‘n_neighbors’ hyperparameter and obtained a final test accuracy of 71%.

Results and Implications

Looking at the final test accuracy scores for all 3 models, we can see that the logistic regression model gave me the best accuracy. At present, I am unsure of why my other two models did not perform as well as my first or what other machine learning models would be better to implement. In the future, adjusting the hyperparameters for each model would be ideal to improve the accuracy and to obtain a more optimized model while avoiding overfitting and underfitting.