


Project 7

Link to presentation slides:  Project 7

Link to GitHub repository: https://github.com/AndrewCurtis27/CS6830_proj7

Introduction

The first dataset contains variables that are related to credit card fraud. The goal of this analysis is to create a classification model to predict whether a purchase is fraudulent, based on certain variables, such as the distance from the cardholder's home or whether it was an online order. In our analysis, we examine which variables are highly related to fraud and test logistic regression and support vector machine algorithms to see which one has the best predictive ability. Fraud detection is increasingly important as social engineers are quickly develop new methods to access credit card information. Our analyses will aid in identifying and stopping fraud.

The primary goal of the second examination is to learn a model to predict whether an individual will earn a salary of more or less than \$50,000 per year. This is done by predicting from information collected from the 1994 census and testing using a logistic and svm model. Some important features for the svm model are education level, age and hours-per-week worked. An additional examination was made to determine whether it would be important to complete a 4-year college degree in order to achieve this income cutoff. This model could be used by the government to determine trends in the economy or by financial institutions to determine financial stability when deciding whether to give out loans or conduct business with an individual.

PART 1: FRAUD

Dataset

The dataset was obtained from [Kaggle](#). It contains one million samples of 8 variables. The target variable is fraud, which is binary, indicating whether it was present (1) or not (0). The numeric predictor variables are the distance from home, distance from the last transaction, and ratio to median purchase price. There are also four binary variables: whether the purchase was from a repeat retailer, whether the chip from a credit card was used, whether the pin number was used, and whether the purchase was an online order. There were no missing values, and relatively little data cleaning needed to be done. However, we did standardize the numeric variables, as they were all on different scales. We opted to take a random sample of 100,000 to use in our analysis to speed up the SVM models. Overall, this dataset was suitable for the analysis.

Analysis Technique

We trained a logistic regression and multiple SVM models on the data. Since this is a binary classification task, both of these models were appropriate. The analysis technique is as follows. First, we standardized the numeric variables, as mentioned above. Then we split the data into a training set and a testing set. Then we examined both the mean absolute deviations (mads) of the predictor variables and the correlation matrix of all the variables in the training set to determine which features could be the most predictive. However, since there were relatively few variables, we opted to keep them all in. Then, we examined scatter plots of some pairs of variables to visualize how instances of fraud were broken out between them. This allowed us to compare the decision boundaries created from the models later on. We

then trained a logistic regression model, a SVM with a polynomial kernel, and an SVM with an rbf kernel. Additionally, we trained two other rbf models to examine differences in class weights. We used 10-fold cross-validation to train the logistic regression model and 5-fold for the SVMs.

Results

Ratio to median purchase price and whether the purchase was an online order had the highest mads (0.53 and 0.45, respectively). This variability means that they could be important predictors. They also had some of the highest correlation coefficients with fraud ($r = 0.46$ and 0.19 , respectively). Distance from home had the next highest values. As such, we plotted the relation between the ratio to median purchase price and the other two variables mentioned in Figures 1 and 2 for comparison later on. Overall, it seems that more online orders tend to be fraud, and the more extreme the ratio to median purchase price and distance from home, the greater chance of fraud.

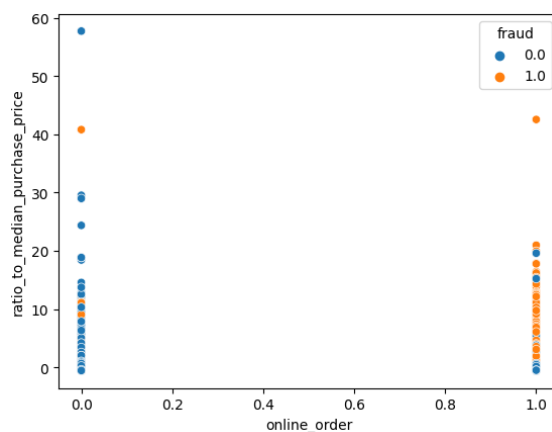


Figure 1

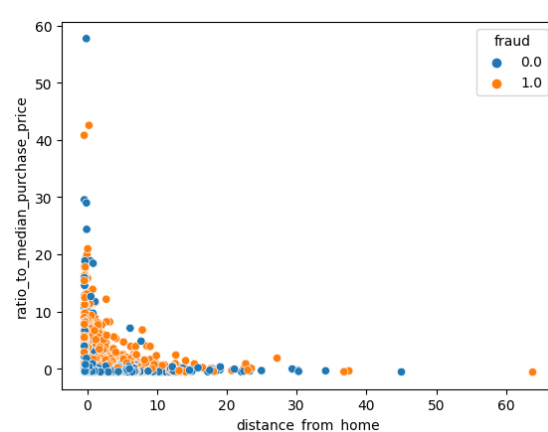


Figure 2

The mean F1 score from the cross-validation for the logistic regression model was 0.71, but the F1 score on the test data was 0.98. This could mean the model may not be as generalizable, or we did something wrong during cross-validation. This model only took 0.19 seconds to run. The decision boundaries for the variable pairs mentioned above can clearly be seen in Figures 3 and 4. Contrary to the original assumption, it looks like only low levels of ratio to the median purchase price (e.g., when a purchase is not far from the median) are not predicted as fraud. Figure four is zoomed in to see the decision boundary clearly.

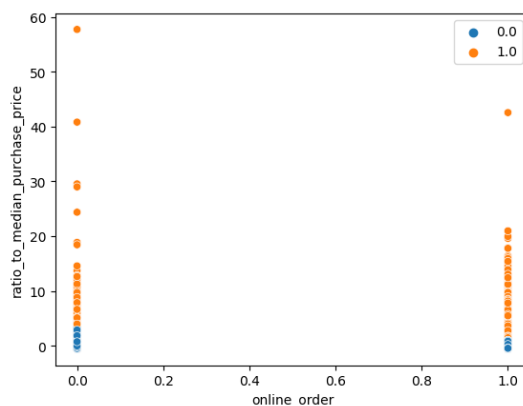


Figure 3

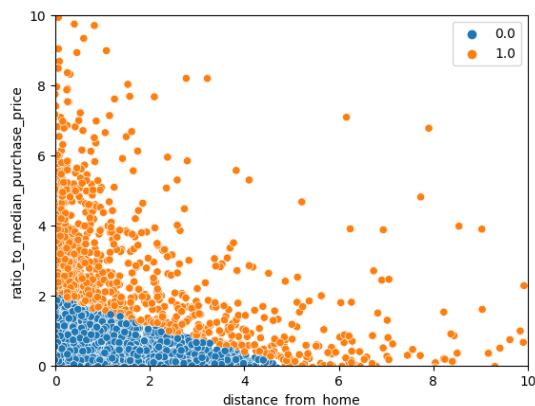


Figure 4

The mean F1 score for the cross validation with the SVM with the polynomial kernel was 0.95, and the F1 score on the test data was 0.995. However, the run time was 32 seconds, which was the longest out of any model. The plots of the decision boundaries for the selected variables are given in Figures 5 and 6. Figure 5 shows how the polynomial kernel is able to give more robust decision boundaries, as all of the non-online orders are not predicted as fraud and the online purchases that are close to the median purchase price, or extremely far away, are not as well. Figure 6 shows that only extreme values are predicted as fraud (note: we cut down the data to make this plots because they took forever to run without it).

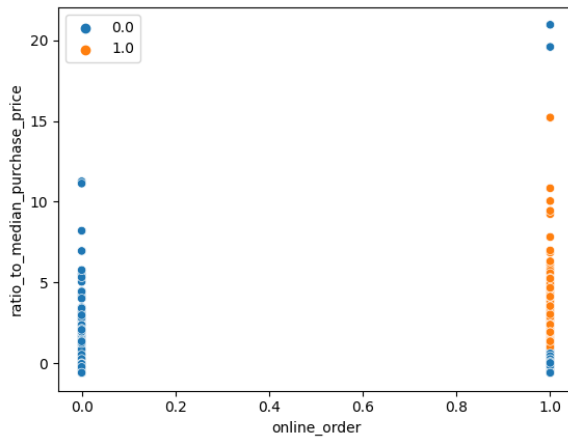


Figure 5

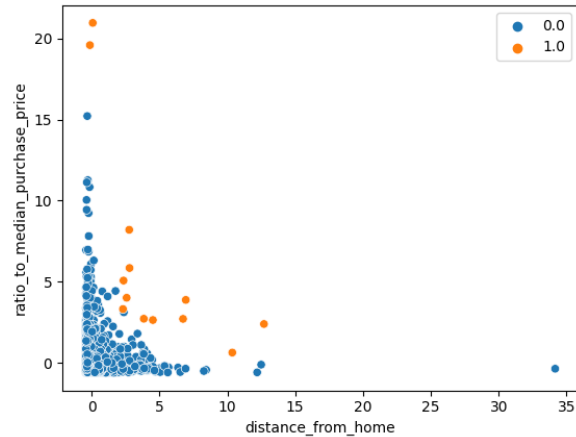


Figure 6

The mean F1 score from the cross validation with the rbf kernel was .97 and the test F1 score was .997. This was the highest performing model. It also only took 13 seconds to run. The decision boundaries of the selected plots are given in Figures 7 and 8. These plots look like somewhat of a mixture between the previous models.

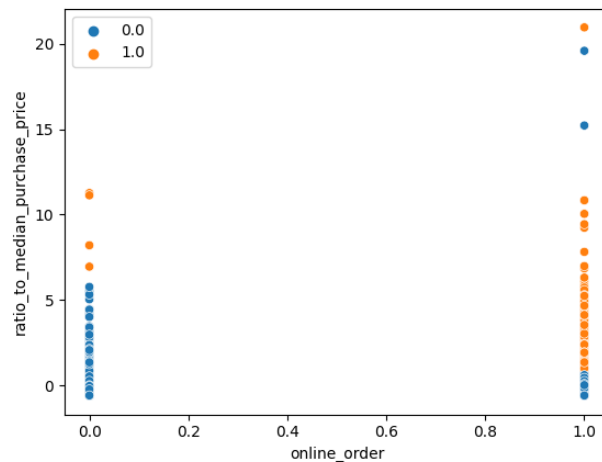


Figure 7

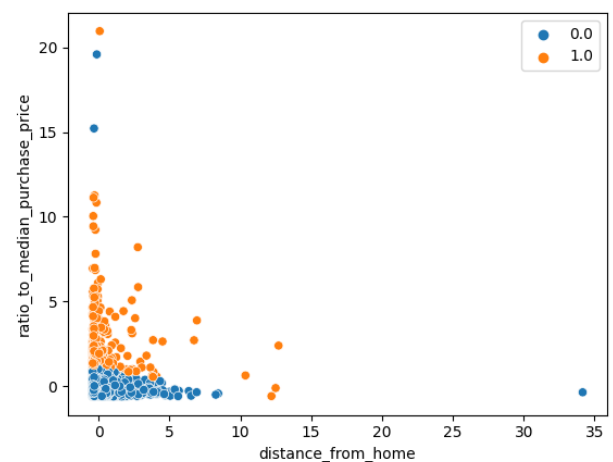


Figure 8

Figure 9 shows the scatterplot of distance from home and ration to median purchase price with the weight of 1 (i.e., fraud) changed from 1 to 5, which means the model is more likely to predict fraud, as can be seen in the figure. Figure 10 shows the converse, when the weight of 0 (i.e., not fraud) is changed from 1 to 5, while fraud is still one. This classifies the majority of the data as not fraud.

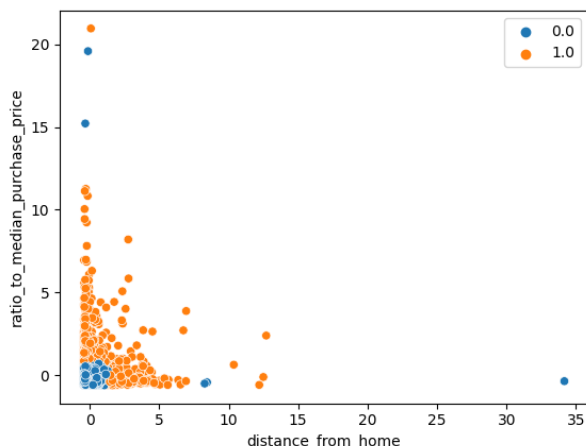


Figure 9

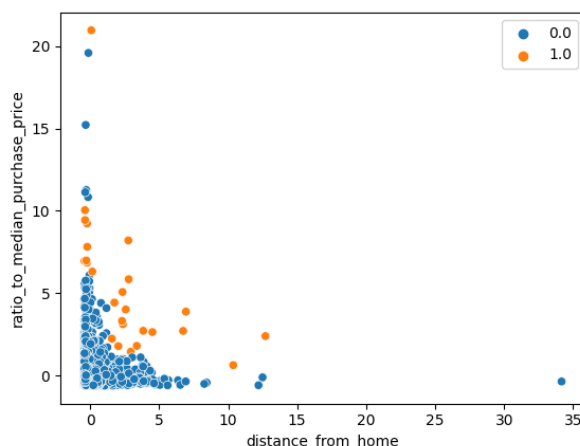


Figure 10

Technical

Many of the technical information is explained above. It seems like the logistic regression or rbf SVM models would be the most generalizeable. The polynomial SVM has such a unique boundary that it may not generalize well to other data. This model also takes significantly longer to run than the other models. In practice, I would use the rbf SVM model because of its high performance. It would also be useful to weight the predictions more towards fraud, because it would be better to flag instances of fraud that were not fraud than to miss instances of fraud. Since the SVM models used in this analysis are OVO models, they take more computation power, and require a longer runtime. Overall, these analysis help show what is more useful in fraud prediction.

PART 2: SALARY PREDICTION

Dataset

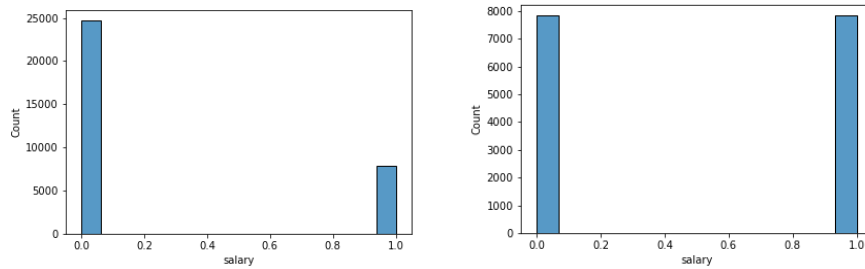
The second dataset used for analysis is the Census Income dataset, also known as “Adult” dataset. It was collected from the 1994 census and includes 32,562 records from individuals over the age of 16. The goal of the dataset was to perform a classification task of whether an individual makes above or below \$50,000 of income a year. The dataset includes many categorical variables, some of which are ordinal and several continuous variables.

Analysis Technique

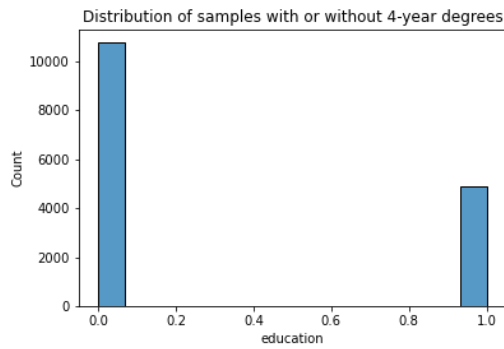
The method of analysis for the Census Income dataset relied first on several data preprocessing decisions. Nominal categorical variables were used for the logistic regression model and ordinal variables were numerically encoded. Specifically, education was encoded 0 for below a 4-year college degree and 1 for a bachelor's degree or above. The data was resampled to balance the final predictor variable class of income being above or below \$50K as it was imbalanced towards below. Many models of a support vector classifier were trained to test for values of the hyperparameter C and kernel type as well as variables to be used and their effect on performance. A logistic classifier was also trained using the nominal categorical variables. The models were trained using an 80/20 train test split and averaged over 10-fold cross validation.

Results

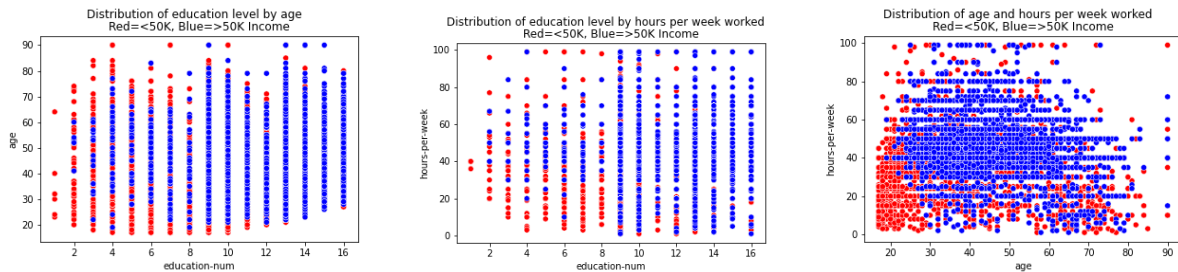
The initial distribution of income for below \$50K (0) and above (1) can be seen below on the left, after resampling the class has been balanced and can be seen on the right, leaving 15682 samples.



After salaries are balanced is there still a difference in education level? This can be seen below, about twice as many individuals do not have 4-year degrees even accounting for income.



Below is an examination of the relationships between suitable continuous variables for svm in the dataset: Age, Education-num, and hours-per-week worked. The plots are colored such that the red points are for salaries below \$50K and blue is above.



Based on these plots a few conclusions can be interpreted. Individuals are much more likely to make below \$50K in income if you did not complete a highschool diploma, regardless of age or hours-per-week worked. It is also highly likely that an individual will make less than \$50K if they are simply under the age of 30. The cutoff of a high school diploma appears to be a more important factor than having a 4-year degree when determining income.

The variables were first tested individually to determine which had the best predictive capability, the results of which are summarized in the table below.

Variable	Average Accuracy	F1-Score
Age	0.64	0.65

Hours-per-week	0.63	0.64
Education-num	0.67	0.66
Education	0.65	0.64

The logistic classification model was trained using the nominal categorical variables. The average accuracy was 0.700 and the F1-score was 0.70. Surprisingly, this model performed worse than the SVM even though it contained a number more variables with seemingly important social information.

A svm model with education-num, age and hours-per-week worked was chosen to be the best set of variables and a tuned-model with hyperparameters and kernel was selected. The chosen value of C was 1 and a polynomial kernel. This produced an average accuracy of 0.726 and an F1-score of 0.74. For context, the best linear model had an average accuracy of 0.720 and F1-score of 0.72.

Technical

For the Census Income dataset, initially models trained without balancing the income class would all return the same value as it would predict only for one class. This was solved by taking a resampling of the income class and balancing it, leaving about 15,000 samples. A dummy variable of education was added to denote if an individual had attained above or below a 4-year degree. The models were trained using an 80/20 train test split and averaged over 10-fold cross validation. For the SVM, the kernel type only changed the performance by about 2-3% and varying the hyperparameter C, affected the models even less.