Michael Haldeman

Credit Card Analysis Porject

This project will be taking in a credit card classification dataset and will be creating models to predict 3 levels of credit score based upon various variables. This could be a very usable model for banks and credit company when needing to approve or deny credit or loans. The model will be trained by 22 different variables that will teach the model how to predict the 3 levels of credit based upon it. Those levels being Poor, Standard, Good. These will be changed to dummy variables for the model to be more efficient in its predictions. The problem that will be solved using this project is to predict the credit score of someone given the different variables that would make up a credit score. To tackle this task, will be to first clean and preprocess the Credit Card data and from there I will take the data set and apply dummy variables to the certain variables that need it. After that the data will be ready to be used in model prediction and evaluation. The predictions will be based on how well the different models will be able to predict one of the three credit scores based on the variables provided within a test set of the data.

Task Definition and Algorithms

The problem to be solved for this data set is to predict the credit score of someone based upon the 22 variables present within the data set. I find this problem to be quite interesting as it considers thousands of data points and creates a challenging dataset to first clean and then use to predict a certain level of credit score based off the data. I think this problem is going to be a challenge, but it will also be a good model in considering many variables and data points and make them useable for a machine learning model. This could be then used for any bank or credit company that could then use the model to approve or deny someone credit based off data provided.

There will be several algorithms and models used to perform the task needed. For algorithms used there are get\_dummies, train\_test\_split, PCA, StandardScaler, make\_pipeline, KFold, Cross evaluation, GridSearchCV, classification\_report, and confusion\_matrix. For Models used there are Logistic Regression, Decision Tree Classifier, and Random Forrest. All the algorithms and models listed above were influential in creating and evaluating the three models created in order to accomplish the task at hand.

With this project I hope to accomplish the problem given within the introduction. As I believe with the given algorithms and models this problem can be accomplished. I believe the models of Logistic Regression, Decision Tree Classifier, and Random Forrest. will behave well with the given data as it is a supervised learning dataset. I think the outcome of this project will be able to predict the credit score level. All three models have the ability to work well with a supervised data set meaning that they work well with data that has predetermined labeling. With this knowledge, I believe that each of the three models will return high prediction and accuracy scores and then it will come down to which of the three performs the best overall.

Experimental Evaluation

The metrics used to evaluate model performance is Kfold and Grid-Search. This to correctly input the right parameters for the model. From this I used a confusion matrix and classification report in order to evaluate further to see how well the model performed.

The process taken in ordering of creating a model and then evaluation was. First to split thr data into test and train sets. From there I was then able to the train data sets and use them to train the three models. From there I created pipelines that included a Standard Scaler and the model itself to then fit the data to the pipeline and from there evaluated the score of each model and the kfold cross evaluation scores. Compared the three models based off each of those metrics and then decided which model performed the best. After the decision, I ran a grid search for the model to find the best parameters for the model and finalize the prediction.

Results of the process above

|  |  |  |  |
| --- | --- | --- | --- |
| Model Type | Logistic Regression | Random Forrest | Decision Tree |
| Models Test Score | 0.673 | 0.684 | 0.723 |
| Cross Validation | [0.673, 0.672,  0.669, 0.657,  0.669] | [0.690, 0.686, 0.691, 0.685, 0.691] | [0.726, 0.716, 0.718, 0.715, 0.728] |

We can see from here that the Decision Tree performs the best out of the three models. From here I took the Decision Tree and performed a grid search to fine tune the parameters for the ability to have the best working model available. In doing so the model performance can then be maxed to the highest it can be.

The grid search included the following parameters and results.

{'criterion': ['gini', 'entropy'], 'max\_depth': [4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 20, 30, 40, 50, 70, 90, 120, 150]}

Best parameters: {'criterion': 'gini', 'max\_depth': 7}

Best cross-validation score: 0.72

Using this I was then able to apply these parameters to the Decision Tree model to obtain the best results.

|  |  |
| --- | --- |
| Decision Tree Test Score | 0.725 |
| Cross-validation scores | [0.724, 0.718, 0.717, 0.718, 0.728 ] |

When compared to the original Decision Tree without a grid search applied there is an increase in the overall score and cross validation scores.

Continuing to evaluate the model we can look to a confusion matrix and classification report.

Chart

Description automatically generated

Table

Description automatically generated

From these graphs and charts, we can first see the number of correctly predicted vs the number of inaccurately predicted data points for each of the 3 variables for credit score. We can see for the most part that the model was accurate in predicting the three different cases. Along with the matrix, we can look at the classification report see what the precision, recall, and f1 score is for the different values. Reading into this, we can see that the values of 1 and 2, being poor and standard have the higher amount of precision. With 3 the good credit having the lowest amount of precision of the three. Overall, the f1 scores of each are on the higher end meaning that they average out to be a relatively good model.

To describe the Decision Tree model. The has a relatively high prediction score when compared to the Regression and Random Forrest. The cross-validation scores help to back up the scores of the Decision Tree model as well. The overall weakness of the model is that since the data is so large and that it is trying to predict one of three values of credit score it be hard for a model to do much better than the 0.73 score that it has gotten on its predictions. This just shows as to how complex the model really is as there are 22 different variables to account for and evaluate to make a prediction upon a person. In the case of the data set being large it also helps to keep the bias down and allows the model to not over generalize the data points as there are many points. The data also needed to be scaled to be workable. If the data was not scaled it would be a much weaker model.

Related Work

For related work based upon a credit card prediction. Credit Karma offers a model simulator that allows the user to input different variables that will then make a prediction of a credit score based upon the inputs. This simulator is a more complex model when compared to the Decision Tree Model formed within this project. The Credit Karma credit simulator is a higher quality model, but the Decision Tree model can be tuned, and the data can be processed better for the model to perform at a higher level. Both models overall do the same thing as it considers different variables and then can predict a credit score based upon the variables.

Future Work

In future works with this data set, I think a different approach with the dummy variables might result in better overall results. As the way this model was coded was that there were dummy variables for the occupation that could have been coded differently and dummy variables for the credit score itself could be improved for prediction. Or if need be, the variable of Poor, Standard, and Good could be changed to be actual values which could then be predicted through the use of the model. The model could then use these predictions and actual values to then evaluate the number value and give it a value of Poor, Standard, or Good.

Conclusion

In summary, I believe that overall, this project was a success. The final Decision Tree model performed well from the data that was given to train it. I believe there are some things that could be done better in terms of data preparation in the future but overall, the model performed well for the problem that was asked. The most important stats that should be repeated are the 0.73 accuracy score on predicting the credit score based upon the data. This is an above average model that can predict the correct score well. Since there are three possible values for the model to predict, for the model to be about 70% correct would be considered a good model. Along with the accuracy score being higher, the F1 score for this model and the three outcomes are all on the higher end. Overall, this model answers the question asked and performs at good level. This project achieves everything that I had hoped for it to do.