Classification Models Project Report

CSC 177 - Section 01 and Section 02

Team: Import TeamName

Team Members: Lauren Prather, Vibhor Sagar, Rachel Mao, Santiago Bermudez, Mueed Khalid, Marco Toro, Jared Roque

Summary:

When it came to working on this project, we first started off by applying the many classification models onto our classification dataset from one of our previous projects, which was a titanic dataset we used for the purposes of predicting whether a certain passenger survived or not. This was all just for practice and to familiarize ourselves with the different models at hand. For part 1, where we were tasked to find the most useful features needed to predict which subscribers to a service will discontinue their subscriptions, we decided to first perform some basic preprocessing on our dataset and then run a logistic model on our features to determine which set of features is most important. Basically, the features with higher absolute regression values are the most important, which led us to conclude that the IsActiveMember, Gender, Balance, and NumOfProducts features are the most important. When it came to deciding on an appropriate train/test split, this decision was more of personal preference since there is no optimal split percentage. We just needed to make sure that the percentage suits the requirements and meets the model's needs. For our dataset, since it had 10,00 rows, we just decided that an 80:20 split was reasonable enough to avoid overfitting and underfitting. Lastly, when it came to running the classification models on our given dataset, we simply did our best to replicate the process we went through when it came to using these models on our own dataset and made changes and corrections as needed to suit the provided dataset.

Results:

ML Model	Accuracy on Test Set (Provide accuracy in %) 0.8535
Naïve Bayes	0.78

KNN	Also provide the K value for which you got the highest accuracy 0.845 Our K value that got the highest accuracy was 7!
SVM	0.7
DT	1
Logit	0.8

Conclusion:

Surprisingly (*or not) enough, we found that the decision tree classification model was the most accurate out of the six we worked with when it came to testing our predictions on the provided dataset. Please note, however, that when it came to making predictions, we used a subset of the original dataset to deal with issues regarding computing and processing power that we have faced when we tried to work on the original dataset of 10,000 rows. After that, our next most accurate model was our machine learning model at 85.35% as opposed to 100% accuracy. Our least accurate model was SVM, followed by Naïve Bayes. With regards to accuracy, we believe that the decision tree only managed to score so well as it had a small dataset to work with. However if we used the original dataset, it is possible that the decision tree's accuracy will be much lower. With machine learning, the machine learning algorithm is often fed with massive amounts of data. Hence, machine learning accuracy dramatically depends on the correctness of this data. Also, machine learning involves building methods that 'learn', which can serve to explain why our ML model performed so well. As for our least accurate model, which was SVM, one thing to keep in mind is that models are often sensitive to parameter optimization. So while our SVM model seems to be underperforming in comparison with the rest of the models, this is only true for the selected parameters and dataset that we have. SVM is a supervised machine learning algorithm which can perform Regression and Classification. It involves using data points that are plotted in n-dimensional space where n is the number of features. While it may not have been good for our dataset and parameters, that does not mean that it does not have its uses. The same goes for all our other classification models.