**CAPSTONE4\_FINAL\_REPORT**

In our objective of determining which factors are best indicative of whether a credit loan applicant would make an ideal candidate for a credit loan or not, a support vector machine or SVM was employed in detecting the class that each applicant would fall under based on these features. With many features included in the file and with the goal of filtering which features are the most impactful in terms of classification of a ‘good applicant’ or ‘bad applicant’, some hypotheses were formulated, and a few features were pre-selected based on compatibility with the hypotheses in addressing the questions posited. Three different hypotheses were formulated in which three different data sets were created and an additional data set that combines all the features of the previous three data sets to get a more holistic view of how all the features together impact the classification of a credit applicant and how the combined features interrelate with one another in forming an applicant’s approval for credit or not.

The first data set created contains a total of five features with four of those features detailing various checking account statuses and where the applicant of each observation or row falls under and the first feature serving as the column containing the labels of good and bad. Feeding this first data set into the SVM object resulted in an accuracy score of 68% correctly classifying each observation based on the features with the checking account values. I postulate that the accuracy score was lower than expected due to the low support for the bad labels with these labels barely reaching above 25% of the total labels of the reserved test size. There simply weren’t enough observations with bad labels as compared to the good labels for the SVM to get better training on to have better prediction abilities for the binary classification. This first data set was created to address the first hypothesis of the amount of funds within an applicant’s checking account having no bearing on if they will be approved for a credit loan nor their class rating affecting it.

Likewise, the second data set was created with a total of five features with the first feature containing the labels for the binary classification of good or bad and the other four features detailing the credit history further broken down from one original feature of credit history. Credit history logs how many credits are paid, and the status of their accounts based on those credits being paid in a timely manner. For this run of the SVM, an accuracy score of 73% was achieved, which is significantly better than the first run. This run of the second data set suffered from the same shortcomings as the first run and likewise for all the other runs that will be explicated on in the following paragraphs. This data set was created to answer the question posited by the second hypothesis which is if the applicant’s credit history would have any impact on whether they would be approved for a credit loan or not.

For the third data set, five features were also implemented with the first feature containing the labels and the remaining four features containing the employment data broken down into four categories from the original feature. An accuracy score of 73% was achieved for this run of the third data set on the SVM object. This data set was created and modified to address the question posited by the third hypothesis regarding employment status; namely, if employment history is one of the decisive factors of whether an applicant gets approved for a credit loan.

The fourth data set created for the final run on the SVM was a combined data set of all three previous data sets to see how each of the features interrelated and influenced which class an applicant gets placed under in the binary classifier. The combined data set contains a total of fifteen features with the first feature containing the labels and the following fourteen features holding the values for the features detailing the checking status, checking history, and employment info. This final run on the SVM achieved an accuracy score of 74% which turned out significantly better than the first run with the first data set and marginally better than the second and third runs with their respective data sets. This run with this data set was not created to address a question posited with a hypothesis in mind but was created for the sole purpose to determine if the SVM would be able to predict each of the classes that the observations belonged to if all the features with the decisive factors were involved.

Regarding potential discussions as to how the accuracy of our binary classifier in the SVM object can be improved upon to achieve at the very least 80% accuracy and as to why four separate runs were conducted, we will delve into that more here. As the SVM is a machine learning algorithm best suited for data sets that are not too large in the number of observations contained and for more modest number of features, I decided to work with this machine learning type as it seemed to be a good fit for the type of data set, we were working with. Although an accuracy score of 80% was hoped for, the 68-74% accuracy score achieved for each of the runs overall was not too great a discrepancy in the actual outcome of running the data sets through the SVM object. Furthermore, as for the kernel type employed in the SVM, out of the three choices of polynomial, gaussian, and sigmoid, the sigmoid should have been the optimal kernel here as it is suited for binary classification problems which our data set just is, but it happened to be the case that the gaussian kernel churned out the best performance and results for our data sets. Perhaps this is due to the data set being a bit complex in its nature especially considering the final data set of combining all three of the previous ones. Also, without accounting for the bad labels for a moment here, the accuracy score would have been well over 80% as the support or number of examples to work with consisted of three quarters of the test size for the ‘good’ labels while the ‘bad’ labels didn’t have much support at all. The original data set itself though did not have a balanced sampling of both labels as 700 of those labels were ‘good’ while the other 300 were ‘bad’. This lack of sufficient support for the ‘bad’ labels is likely what brought the accuracy score lower than our 80% accuracy target as the SVM simply did not have enough of the ‘bad’ labels and observations in which to train on. A data set with more balance in its classifications would be of great help here to improve upon the SVM’s accuracy score here. Increasing the test size to 30% instead of 20% of the original data sets wouldn’t have made a substantial difference here most likely as originally deliberated upon to boost the performance of the SVM.

In addition, the four different data sets and four separate runs for each of those data sets were conducted in the hopes of determining how each of those three original features apart from the feature with the labels were instrumental in swaying the performance of the SVM in training on those particular data sets and thereby getting a reading on the accuracy score and other metrics produced solely on each major feature in focus. Following that, all the features were then combined as in reality each credit applicant would be classified based on where they fall under in all those categories and to also analyze how all those features together influence and impact the performance of the SVM in its performance on the data to see if there is an increase or decrease in the accuracy score and other metrics. These three features were purposely chosen with the others neglected as these features were deliberated as perhaps carrying the most weight in classifying each of the applicants as well as meeting the criteria for the SVM’s optimal and efficient performance in being features containing numeric values which is necessary for the features which hold the values to be predicted on to output a given label. Of course, any other feature so long as it meets the SVM’s requirements could be substituted instead to perform a run on the SVM object if another dimension to the credit applicant’s classification is wished to be analyzed. Likewise, any additional feature can be added to the combined data set with the caveat that one must be cautious with the number of features one uses that holds the values to be predicted on as an excessive number of features is not suited for the SVM and thus require another type of machine learning algorithm to be used.

From both modifying and customizing the original data set to create new data sets that feature specified columns with the categories and values desired, it can give us a clear view of the features that hold the most weight in the various aspects that can affect a credit applicant’s approval for a loan. A careful deliberation and selection of the feature/s can help to reveal and inform us of the variables most in need to be considered and perhaps even shed some features if they appear to serve no useful measure in our classification of applicants. To be noted here however, is that the data sets created were one-hot encoded which entails the conversion of categorial data into a numeric format for the SVM to be able to work with in training and prediction. Another option instead would have been to simply leave all the values under each of their respective original column and performed label encoding instead which converts the categorical data into numeric data as well but does not dissect the original column into a multiplicity of columns for each of the original values of the original column. This method could provide a different lens and exhibit a different dimension to the data as far as its strategic implications are concerned. The way the data is manipulated as far as its features and how those features are implemented and then fed into the SVM may provide a vast array of insights only limited by the user’s needs and visionary inclinations.

As presented earlier in this report, the features fed into the SVM holds much import for the direction and strategy of the business as the features fed with their outputted results in metric readings of value can guide in future business decisions. The insight provided by each feature or any combination of features within a data set can also siphon out applicants that would qualify for a credit loan versus those who don’t. An additional precaution here would be to collect observations that are most evenly and adequately distributed for our binary classification as this would give the SVM more sufficient data to train and predict on to increase its effectiveness and quality of work. With this tool of the SVM at hand, features that are determined to be of little use or importance can be removed from the data set for the SVM to work with more precision and simultaneously tailor the data set to include only those features that hold significant import and features that matter to the business in classifying their applicants.