Project 7 - Logistic Regression and SVMs

Presentation: Project 7
Git Repo: https://github.com/Tyler-Johnston/cs5830-project7

Introduction

For our project on Logistic Regression and SVMs we wanted to look into the purchasing intentions of online shoppers and the churn rate of bank customers. The first dataset touched on information to indicate whether an online shopper's intent to make a purchase or not. The second dataset, on churn rate of bank customers, looks into whether or not a customer of a bank will either stay with that bank or leave that bank. This could be informative to banks by letting them know what factors will lead to a customer leaving, and providing them with predictive measures by understanding their customer base. Our first dataset included features gathered from Google analytics such as the "Bounce Rate", "Exit Rate" and "Page Value" along with features such as Month, Weekend, and Visitor Type. Our second dataset included features such as the person's age, account balance, credit score, and other information that could be used in determining their bank churn status. For both datasets, we used Logistic Regression and Support Vector Machines to perform our analysis. This resulted in models that had decent predicting power.

Dataset (1)

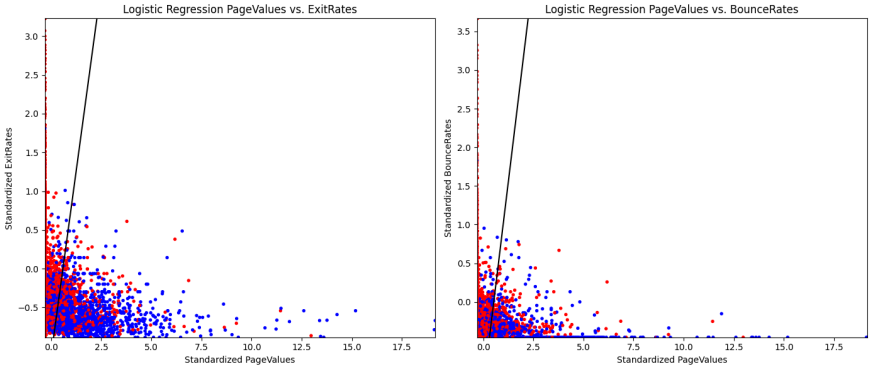
The first dataset we employed was an Online Shoppers Purchasing Intention Dataset. This dataset is used to indicate an online shopper's intention to make a purchase (True or False) using the target variable 'Revenue'. It contained categorical columns such as Month, VisitorType, Weekend, as well as numeric columns such as Administrative, Administrative_Duration, Informational, ProductRelated, BounceRates, ExitRates, PageValues. This information relates to the domain because it provides crucial information that may or may not indicate whether an online shopper is inclined to make an online purchase.

Analysis technique (1)

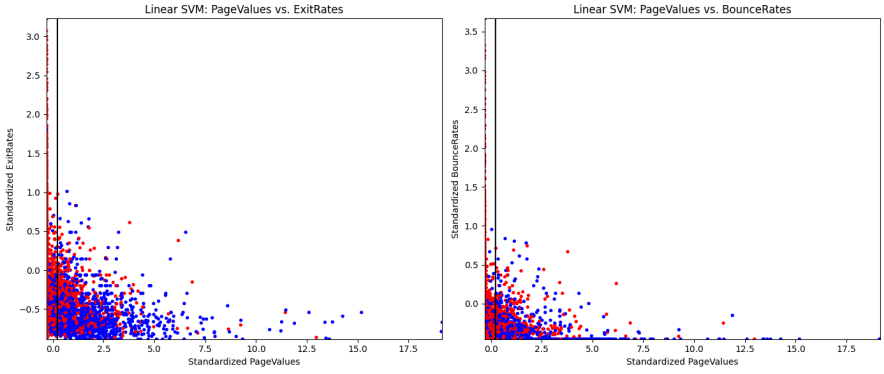
Both logistic regression and support vector machines were employed to predict our target variable of 'Revenue'. However, to determine which columns were the best attributes to be considered in this analysis, a correlation matrix was performed on each attribute against our 'Revenue' target variable. Following this, the correlation values discovered were printed and sorted. PageValues, ExitRates, and BounceRates were considered in this analysis as they were deemed to be very significant in predicting 'Revenue'. To train the logistic regression and SVM models, ExitRates and BounceRates were both compared against PageValues and evaluated against the target, Revenue. For the SVM model, linear, polynomial, and RBF models were also considered. In the case of polynomial and RBF SVF models, cross-validation with train_test_split was employed.

Results (1)

For the logistic regression model, PageValues vs. ExitRates produced F-Scores of [0.49, 0.93], and PageValues vs. BounceRates produced F-Scores of [0.49, 0.93]. However, upon adding a class_weight of 0:0.25, 1:0.75, the F-Scores improved to [0.63, 0.94] and [0.62, 0.94] respectively. This indicates that logistic regression was better at identifying an online shopper's positive purchasing intent rather than the other way around when utilizing PageValues, ExitRates, and BounceRates respectively. Additionally, this same idea was applied to the SVM model. The SVM F-Scores with Linear Kernel for PageValues vs. ExitRates were the following: [0.53, 0.94]; in contrast, the SVM Metrics with Linear Kernel for PageValues vs. BounceRates were the following: [0.53, 0.94]. The training set for Polynomial SVM had the following F-Scores: [0.49, 0.94] while the testing set for Polynomial SVM had the following F-Scores: [0.46, 0.93]. In contrast, the training set for the RBF SVM model had the following: [0.61, 0.94] while the testing set for the RBF SVM model [0.59, 0.94]. This indicates the RBF SVM model showed the best overall results for both positive and negative purchaser intent. Upon adding class_weight={0:0.25, 1:0.75} to each SVM model, F-Scores typically increased. In the same order as previously mentioned, F-Scores changed to [0.55, 0.94], [0.65, .93], [0.56, 0.94], and [0.66, 0.93].



red = absence of purchase. blue = purchase



Technical (1)

Upon loading the shopping dataset, the target variable was converted from boolean to int (either 0 or 1 respectively). Because PageValues had the highest most significant correlation with our

target variable, it was decided to use it against each subsequently chosen attribute. BounceRates and ExitRates had comparable moderately significant correlations with the target variable; thus, these were also chosen for our analysis. Other attributes were tested (they were chosen by looking into the correlation coefficients again), and this process of utilizing other attributes did not yield any significantly different/better results.

Dataset (2)

The second dataset we explored was on bank customer churn. This dataset looks into whether or not a customer will leave their current bank (churn, indicated by 0 for staying and 1 for leaving) and the possible factors that could play into a customer making this decision. Some of the features included are the balance on the account, their estimated salary, age, credit score, country of residence, and a few other features. Many of the features are quantitative, however there are a few that are categorical such as the product number and country. While analyzing this dataset, we chose to use the customers credit score, gender, age, balance, country of residence, and whether or not they were an active member. These features are important to the dataset because they are factors that play into whether a person will leave or stay with a bank.

Analysis technique (2)

While analyzing this dataset, we first had to clean the data and get it set up for the analysis. For this, we One Hot Encoded the country feature, we also had to convert the gender to be more quantitative, as previously it had been either Female or Male, we changed this to 0 and 1 respectively. During the analysis itself, we first looked at how the numeric features correlated with the churn status, this showed us that the age, balance, gender, whether or not they were an active member, and their credit score had a stronger correlation out of each of the features. We then used the Standard Scalar on the features that we decided to use, including the country of residence, in order to standardize our data and have a more normal distribution.
Following this preprocessing, we then performed a Logistic Regression model, and three different Support Vector Machine (SVM) models (Linear, Polynomial, and RBF). Each of these models was suitable for the data because we were using quantitative data to predict a categorical feature.

Results (2)

During the analysis, we decided to focus our model on two features for our graph due to their more prevalent decision boundary already present in the dataset. These features being the customers age and the balance on their account (See Figure 1). As shown in the graph, a majority of the people who left the bank (blue dots) are at a more middle aged age range, whereas those who stay with the bank tended to be either younger or older. Of course there is quite a bit of crossover where we see that there are a lot of middle aged people who did stay with the bank. This made finding a decision boundary a bit more difficult.
When we created our Logistic Regression model for this dataset, we were able to get an f-score of .49 for predicting the person would leave and .83 for predicting the person would stay, precision of .42 and .88, and recall of .59 and .79 respectively. This would indicate that the model was better at predicting that the person would stay rather than leave. As shown in Figure

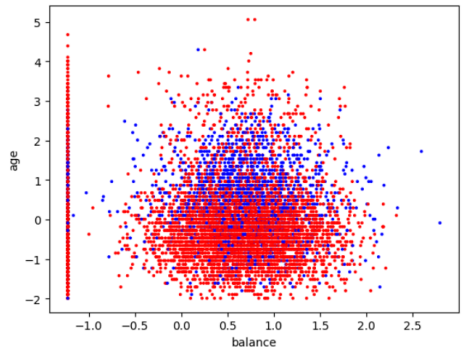


Figure 1: Visualization of age and account balance, blue dots = left bank, red dots = stayed with bank

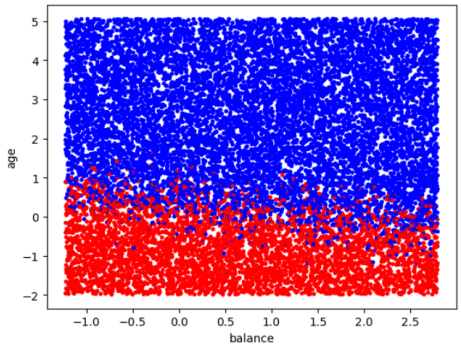


Figure 2: Visualization of Logistic Regression model predicted decision boundary based on 15000 random uniform points

2 compared to Figure 1, the model would not have been able to predict the cases of staying if the person was in the older age group.

We then looked into using the SVM model to predict the churn status. We found that the Linear SVM model was quite similar to the Logistic Regression model (See Figure 3 Graph 1), however the Linear SVM model performed slightly better with an f-score of .50 and .83, precision of .42 and .89, and recall of .64 and .77 respectively.

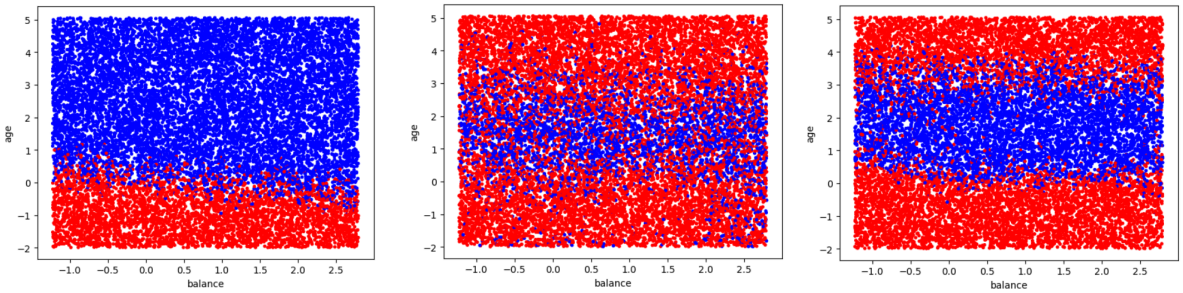


Figure 3: Visualizations of each SVM model and their performance creating a decision boundary. Graph 1 (Left): Linear, Graph 2 (Middle): Polynomial (degree 3), Graph 3 (Right): RBF.

For the polynomial SVM, we decided to set the degree to 3 to hopefully get better results, and we were able to get a clearer decision boundary that included both the younger and older customers' trend in staying with the bank (See Figure 3 Graph 2). For this model, we were able to get an f-score of .52 and .82, precision of .42 and .90, and recall of .68 and .76 respectively. Finally, we tried the RBF SVM model which gave us our best results, providing a very distinct decision boundary and f-score of .55 and .87, precision of .52 and .89, and recall of ,58 and .86 respectively.

Technical (2)

As stated previously, we had to One Hot Encode a few features to make them suitable for training. We also had to use the Standard Scalar to normalize the data. Besides doing that, the data was clean and ready for the analysis.

For our analysis techniques, we used Logistic Regression and Support Vector Machines because this data seemed best suited for those models due to its somewhat distinct decision boundaries, its quantitative features, and its categorical target feature.
During our analysis process, we looked at many graphs to determine which features had the most distinct decision boundary already in place to visualize while creating our model, this happened to be the age and balance features. However, when creating our model, we included a few other features to hopefully improve its ability to determine a decision boundary. We had to include the class weights in each of our models because without it, the model would always end up predicting the person staying with their bank. That could have been a failure on the part of the dataset we decided to use.