CS5380 Project 7: Logistic Regression & SVM

Zach Jordan and Samuel McMillan

Introduction

Our datasets consist of three very differing topics, diabetes, hotel bookings, and red wine. Diabetes is one of the world's most common endocrine disorders, it can cost people their lives and thousands of dollars trying to live with its side effects. This means early prediction and diagnoses can be extremely beneficial to those who may come to suffer from it. Our dataset includes indicators and standards set by the World Health Organization, our data spans over 1304 different samples. With our analysis of these indicators and standards, we hope to develop a model, using logistic regression and SVMs, that can accurately predict the development of diabetes in patients.

Hotel bookings happen all the time. Hotels also deal with many people canceling their bookings, so hotels can lose lots of money by not charging enough on insurance for people canceling at the last minute. That said, it is extremely important to be able to predict how often and maybe even who will cancel, so hotel managers can adjust their insurance prices accordingly. Our dataset includes over 100000 samples from a city hotel and a resort hotel. With the use of logistic regression and SVMs, we hope to be able to find what features and what things lead hotel bookings to fall through for better insurance price estimates for employers.

Red wine is something enjoyed by millions around the globe, it is something many people like to consume but oft they want to branch out and try new wines. The dataset includes many physicochemical (physical and chemical) properties of different red wines and their quality. Through our analysis of these different properties, we can achieve a logistic regression & SVM model that can accurately categorize good and bad wines. Doing so can help wine enthusiasts around the globe further their exploration of the hobby and industry safely without having to waste money on poorly-tasting wines.

Link To Github

Link To The Slides

Dataset 1: Diabetes Diagnoses

Description:

Diving into the diabetes dataset further, it consists of 1304 samples of patients who tested positive for diabetes. The features of the data collected from said patients cover their age, gender, BMI, blood pressure, glucose levels, lipid profiles, liver and kidney function markers, smoking and drinking status, and family history of diabetes. We aim to use these different features and find a way to fine-tune them so that we can accurately predict the development of diabetes

Methods Used:

We employed logistic regressions as our analysis technique for predicting diabetes. Logistic regression offers greatinterpretability which helped us out. Before we got into using our analysis technique we had to dive into the nitty gritty and work on data preprocessing, and feature selection. For our feature selection, we used logit regression to calculate p values. With those calculated p-values we kept statistically significant features, those with p-values less than 0.05. To help make sure our results were accurate we used cross-validation to verify.

Results:

We first set a baseline and ran our dataset through a logistic regression model using cross-validation. Our initial results were quite promising, our mean precision score was 0.95 with a recall of 0.89 and a F1 score of 0.92. This did scare us a little because we may have some overfitting going on, yet we pressed on. After doing feature selection based on p-values we ran it through cross-validation again and got 1.0 for precision, recall, and f1-score. This seems like overfitting and it would be cool to see how it ultimately performs outside of this dataset. In the end, we feel like our model can at least be a starting point in helping catch diabetes early and hopefully save people some money and even more so their lives.

Dataset 2: Red Wine Quality

Description:

The red wine dataset consists of win characteristics. These include aspects of acidity, citric acid levels, chlorides, free sulfur, the pH of it, sulfides, etc. With the use of these features, we hope to be able to predict the quality of wine. The quality ranges from 0-10, but for purposes of logistic regression and SVM, we will consider good to be 6.5-10 and bad to be 0-6.5.

Methods Used:

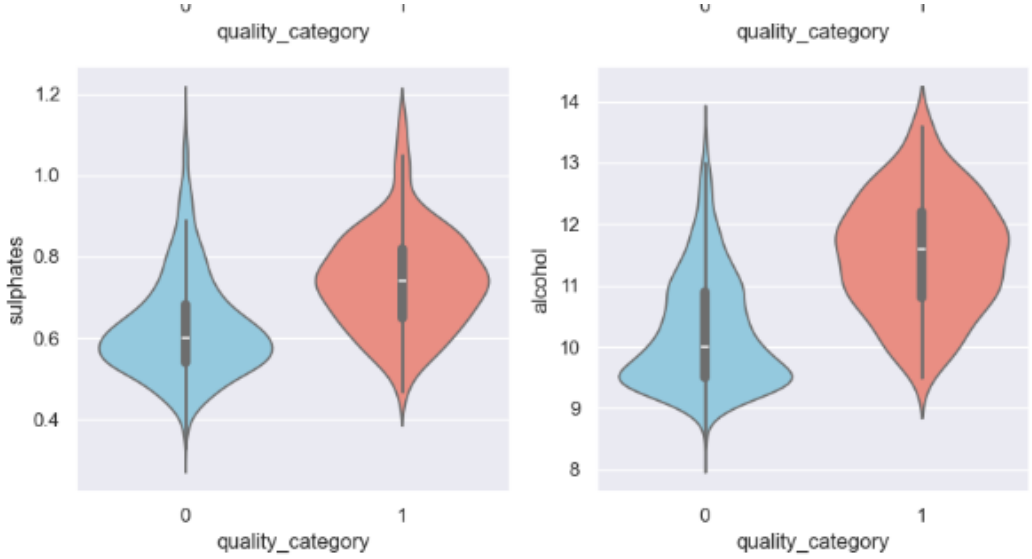
Initially, the logistic and SVM models were trained using each of the features within the dataset. From there, using z_scores remove any outliers that may be making the results skewed. Using heat maps and violin plots, look at features individually and collectively to see

which features to remove or modify. From there, use the most correlated features to the target feature (i.e., wine quality), and make predictions from that.

Results:

```
Logistic Regression Baseline Accuracy: 0.859375
SVM Baseline Accuracy: 0.85625
```

The accuracy for both Logistic Regression and SVM is the without any feature selection



The above shows two violin graphs comparing features to the quality category (i.e., this was done on every feature but only two were chosen to not be redundant to the viewer). In this example, we can see there is a bigger correlation between alcohol and sulfates when it comes to predicting the quality of the wine.

```
Logistic Regression Accuracy: 0.8938356164383562
SVM Accuracy: 0.9006849315068494
```

The above accuracy scores for logistic regression and SVM was from using just alcohol, citric acid, and sulfates as the predicting variables. As can be shown, this is much more accurate than the base prediction using all the features showing that it is only necessary to sometimes look at a few factors to figure out wine quality!

Dataset 3: Hotel Booking

Description:

The hotel dataset consists of booking information for city hotels and resort hotels, this includes things like when the booking was made, length of stay, the number of adults, children, and/or babies, as well as the number of parking spaces among other things. The dataset included thirty-two different features. With the use of all these features, we aim to create a

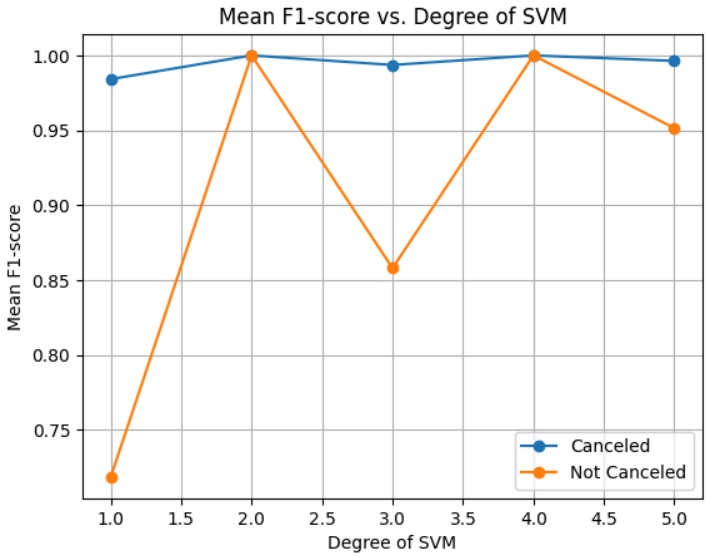
model able to predict which customers are going to cancel and help hotels manage that accordingly.

Methods Used:

The methods used for this analysis were very similar to the diabetes diagnosis analysis. We used cross-validation to get a baseline for our model. Afterward, we worked on feature selection, like the diabetes diagnosis analysis, we used p-values but this time we kept only the five most significant features. We chose to do it this way because of the large amount of different features. Another thing we did differently in this analysis from the diabetes analysis was we employed both logistic regression and support vector machines as our analysis techniques.

Results:

The baseline for logistic regression using all thirty features gave us mean f1 scores in the range of about .39 to .52, so we had room to improve and train our model. After choosing the five most significant features reservation status, country, assigned room type, market segment, arrival date day of the month we ran it through cross-validation again. With this we saw the f1 score to stay about the same range, .26 to .6. It's possible the features we selected were done so poorly, but we wanted to test it out with SVMs as well.



As you can see above we saw much better results with the SVM and our f1 score started to approach higher scores of around .99 and even 1. Once again the near-perfect f1 scores may point to overfitting so it would be cool to test this on other hotels' data and see how our model performs. Overall our analysis has proven successful and our model can be used by hotels to help manage guests who cancel.