Project 7

Nithya Alavala and James Sanford-Luevano

Introduction

The objective of the following project report is to use statistical analysis and data science techniques to access United States housing data and mushroom data. The two major techniques we utilized in this project were logistic regression and SVM(Support Vector Machines). We choose these data sets for a number of reasons, the first of which is we wanted to highlight the utility and versatility of our analysis techniques, and we felt that by choosing such distinct datasets we could show how powerful these analysis techniques were. In our housing data analysis we made predictions on how affordable housing is across the country, and in our mushroom data set we made predictions on the toxicity of mushrooms. Our findings would be valuable to a wide range of individuals, from data scientists that would like to utilize the analysis techniques mentioned above, to those interested in socioeconomic impact of housing affordability, to people interested in eating non-toxic mushrooms and for the food and packaging industry to identify poisonous mushrooms. The presentation can be found here: Presentation and the github can be found here: Github.

Dataset

We worked with two very distinct data sets for our project, housing data in the United states and mushroom classification data. The housing data had information regarding the price of a home, its address number of rooms, even the average income of individuals in the surrounding area. In this data set we decided to use both the logistic regression and SVM methods of analysis. Our mushroom classification data held information regarding the mushroom appearance, population, and habitat. This dataset is obtained from UCI (Unique Client Identifier) Repository of Machine Learning Database.It includes description of the hypothetical sample which is corresponding to the 23 species of the gilled mushroom in the Lepiota and Agarics Family. Each one of those species is identified as the definitely edible, definitely poisonous, unknown edibility, or is not recommended at all. The latter class has been combined with a poisonous and edible based on 22 physical attributes. We decided on using the logistic regression and SVM analysis techniques for these datasets.

Analysis Technique

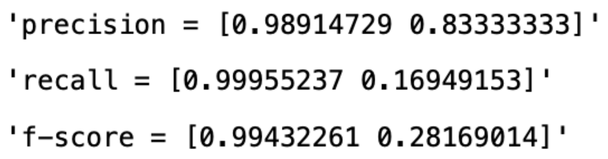
For our housing data set we decided to analyze the data from a socioeconomic perspective. We wanted to predict if the people within a given community could afford a home in their community. For the analysis we decided to take the price of the home and the size of the surrounding community and see if we could predict if a person in said community could afford a home in it.

We utilized logistic regression and SVM(Support Vector Machines) techniques while analyzing this data set. We also attempted to see if we could use the price, population, and average income of the surrounding community to tell which state the home was located in; we received inconclusive results when attempting this.

For our mushroom data set we decided to analyze the data from a toxicology perspective. Our analysis involved predicting the mushrooms if they are edible or poisonous which is a binary classification problem. We decided to include all the features from our data that contribute to satisfy certain insights such as only edible mushrooms have rooted stalk root, only poisonous mushrooms have buff or green gill color and other similar insights which helps to indicate whether a mushroom is edible or poisonous. For our analysis, we first did some data preprocessing by encoding categorical variables using label encoder and then implemented one hot encoding on those variables and performed data cleaning. In our analysis of the dataset, we applied both logistic regression and SVM approaches.

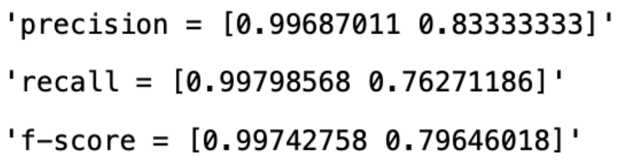
Results

When comparing the two methods with respect to the housing data we found the results were mixed, the following table show the metrics of the logistic regression model:



```
'precision = [0.98914729 0.83333333]'
'recall = [0.99955237 0.16949153]'
'f-score = [0.99432261 0.28169014]'
```

The results were skewed in that we had better predictive capability in instances where the average income of a community did not meet the minimum salary to purchase a home and this could be attributed to the fact that it was more prevalent than the average income of a community meeting the minimum income to afford a home.



```
'precision = [0.99687011 0.83333333]'
'recall = [0.99798568 0.76271186]'
'f-score = [0.99742758 0.79646018]'
```

However if we compare the results from our logistical regression model to the SVM method of analysis we see that the SVM significantly outperforms our previously applied method.

Model	F1 score
Logistic Regression	0.9943
Linear SVM	0.9974
Poly SVM	0.9968
RBF SVM	0.9972

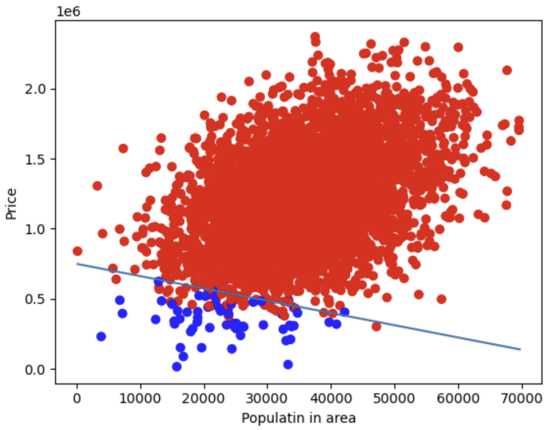
While comparing the methods with respect to the mushroom dataset, we observed the results mentioned in the table below. If we compare the results, we can say that the accuracy of RBF SVM is significantly higher compared to other methods used on the dataset and it has been found to outperform these other methods in predicting the mushrooms.

Model	Accuracy
LR	0.9635
LR using GridSearch	0.9630
Linear SVM	0.9699
RBF SVM	0.999
Poly SVM	0.998

Technical

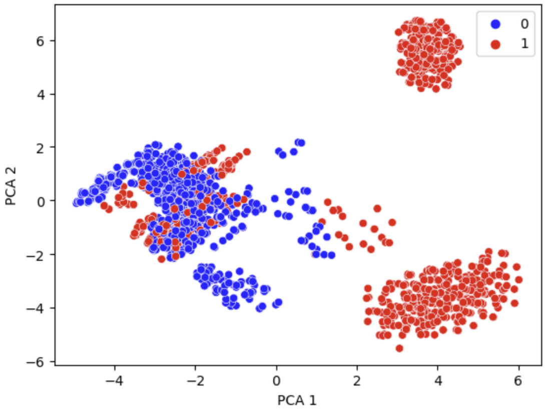
We found that for our housing dataset the SVM predicts well when average population and housing price. We found that both methods predicted well but SVM outperformed logistic regression.

Housing data plot:



We found a slight difference between our polynomial and RBF SVMs. We found that in general the RBF performed slightly better.
While running the different models we found the run time of the SVM to be significantly longer than our logistic regression model.

We did some data preprocessing on our mushroom dataset by encoding categorical variables using label encoder and then implemented one hot encoding on those variables. And then we found that this feature, 'veil-type' has 0 values and was not contributing to the data so we dropped it and checked the correlation of all other features to drop highly correlated ones. Additionally we standardized our data and applied PCA to the features in the dataset to reduce their dimensionality and used it to transform the original data set into a set of 10 principal components. To visualize this, we plotted a scatterplot with the first two principal components to understand how well the transformed features separate the two classes.



We observed that for our mushroom dataset, we first fit the data to our logistic regression model and then used gridSearch to get the best parameters and compared these two models to predict the mushrooms but no difference in accuracy, it achieved around 0.963 where we noticed that even with best parameters, it did not show any difference in performance for logistic regression. And then we also attempted to try changing class weights by giving a range of values, and found that accuracy was higher for 0.4 and 0.6 as the class weights for 0 and 1 respectively compared to other values. While there was not much effect on the data when changing the class weights in SVM. SVM predicts well with accuracy of 0.99. We found a slight difference between the polynomial and RBF SVMs. And we are likely to say that SVM outperformed logistic regression. SVM model took considerably more time to run compared to our logistic regression model.