Seth Beckett & Dave Story
CS 6830 Project 7
Links: Slides and GitHub

Introduction

We have explored two different datasets in the hope of deriving useful information. We chose to focus on the air traffic industry and the biological sector (mainly mushrooms). Our air traffic analysis will help airlines satisfy their patrons, helping them maintain customers and increase profit. Our mushroom edibility analysis will provide insightful info for outdoorsy folk who are interested in collecting and eating wild mushrooms.

Airline Satisfaction

Dataset:

The airline dataset consists of multiple features that can help predict if a passenger will be satisfied with their trip. Most of the data is in the form of rankings 1-5, however there is data on wait times and travel distance. It contains an overall satisfaction column, in which passengers reported if they were satisfied in general with their trip. It is this column that we hope to predict.

Analysis Technique:

We wanted to use a support vector machine (SVM) to predict if a passenger was dissatisfied with their trip. We also used scatter plots to visualize decision boundaries for our models, this would help us fine tune our SVM. Heatmaps were used to find important attributes that predicted the satisfaction level of a passenger.

Results

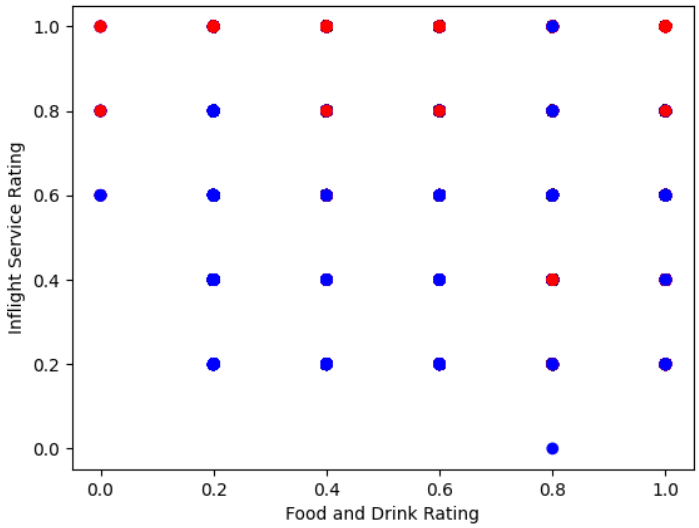
We tried different SVM strategies when building our model. They each had different advantages, but it will be easiest to compare their predictive capabilities.

Model Strategy	Precision	Recall	F1
Logistic Regression	NaN	NaN	0.87
Linear SVM (0:4) Weight	0.92	0.76	0.83
Poly SVM (degree 2)	0.93	0.96	0.94
RBF SVM	0.92	0.96	0.94

We see that logistic regression performed the worst, while a poly SVM and RBF SVM performed about the same. However we did find that RBF SVM takes about twice as long as the poly SVM, so if computation time is important that may influence our decision onward.

After finding a good SVM strategy, we began to look at the most impactful features our model was using to predict if someone was dissatisfied with their trip. Most of these features were out of the control of airlines. However food, drink, and inflight service were features that were important features that airlines could control.

The following chart shows the relationship between these two features. Red is satisfied, blue is dissatisfied.



We can infer a decision boundary using this data. We hope that airlines can use this data to better serve their passengers, and therefore become a more profitable airline.

Technical:

Data preparation wasn't too hard. We did need to one hot encode some columns, but most of the preparation came from normalizing. Most of the data was on the scale of 1-5, but we had very large numbers like flight distance to worry about. Normalizing fixed these issues. We ultimately went with the RBF SVM because of how much data we had, and how many features we used. There was so much data that we ended up throwing away 70% of the dataset for each training run to reduce computation time. This is fine, as we still had 30,000 samples to work with by doing this. The process to find our best features was first attempted by using a heat map. That didn't yield anything noteworthy, so we moved on to training a model and getting its most favored features.

Mushroom Edibility

Dataset:

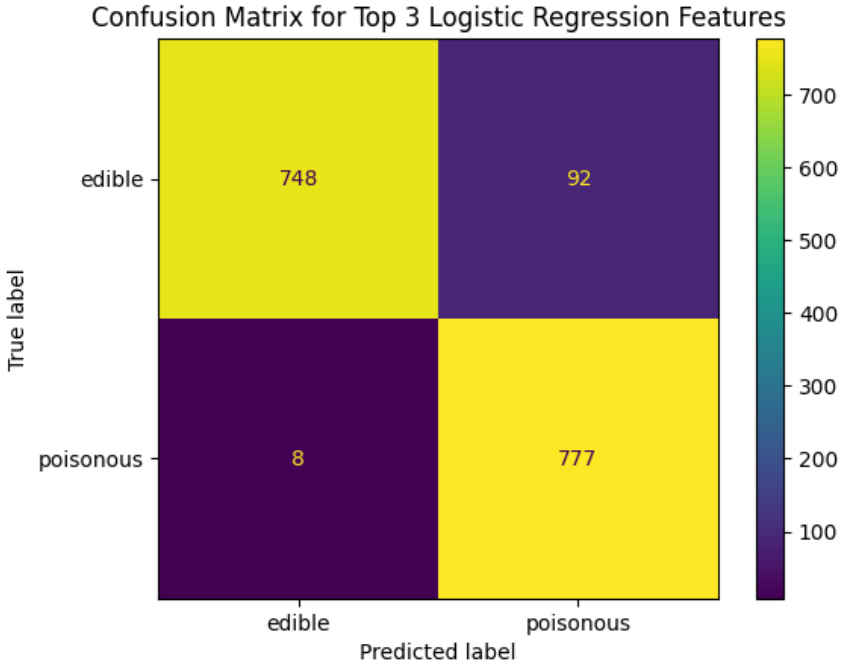
The mushroom dataset contains around 8000 observations of different mushrooms' physical characteristics and whether they are edible or poisonous. Characteristics included attributes about the mushroom cap, mushroom gills, and mushroom odor. All features were categorical. Our target attribute for prediction was edibility.

Analysis Technique:

We tested using both support vector machines and logistic regression to predict mushroom edibility. We found that both worked exceptionally well, but that when we wanted to create a model with the least amount of predicting features possible, logistic regression outperformed support vector machines. We ended up training a model using three categorical features: absence of odor, anise-like odor, and green spore prints or not.

Results:

Our mushroom results are that you can safely predict whether or not you can eat a mushroom 94% of the time, solely based on the three aforementioned attributes. If the mushrooms have no odor, or an anise-like odor, that's a good sign for edibility. However, if they have a green spore print, you definitely do not want to eat those. The following confusion matrix shows how our model performed on test data:



Technical:

The data prep for the mushroom dataset was very simple, since everything was categorical we just made dummy variables for every category. Support vector machines and logistic regression turned out to both perform amazingly well for this task. Originally, we used a grid search to cover all kernels of SVMs and to test a wide variety of parameters for both logistic regression and SVMs. We found that when using all of our features, even a very simple linear SVM or logistic regression model would predict with 100% accuracy.

We then explored if we could get high accuracy by drastically cutting back on the amount of features used to train our model. Since coefficients are directly interpretable with logistic regression and linear SVMs, we were able to select the top coefficients by sorting in order of highest absolute value of the coefficients for our full models fitted, and then select the top features from there. We found that you can still get 99.5% accuracy with just 5 features, whether the model is a linear SVM or logistic regression. Furthermore, the logistic regression model with only 3 features still would very accurately predict whether a mushroom was edible (as seen in the above confusion matrix), so this was the final model we decided to use. The linear SVM with the top three features had poor accuracy, precision, and recall.