INFO 1998 Project C Deliverable

Logistic Regression, SVM, and Decision Trees

Guideline and Rubric

Release Date: March 21st
Due Date: April 10th at 11:59 pm (Note this is 3 weeks so it isn't due over spring break, but part D will be due April 17th so plan accordingly)
Submit Through: CMS

Overview

We have now finished learning about the many different types of supervised models in lecture. For this project we are focusing specifically on logistic regression, SVM, and decision trees. You will be using the same train.csv from the housing dataset for this project as well. We again encourage you to use **your modified csv from part A**. In addition to learning how to use the new models, feature engineering remains an important portion of this project. So be aware of how you are selecting your features to maximize accuracy.

As a quick recap, logistic regression is used for binary classification. SVM is another classification model based upon finding the best hyperplane to separate data. Decision trees can be used for both classification and regression and splits the data in a tree-like shape. The baseline for classification models is calculated as the percentage of the column's most frequent value. Remember, model improvement is a function of the baseline. If a percent improvement seems very large, the baseline is probably very small.

For all of your models, there will be no restrictions on the columns used as features (except for the logistic regression model) and feature engineering is encouraged. Same as before, you should create **five different train/ test splits** that all have the specified percentage improvement over the baseline. You should be using random_state to set the seed for random splits so your data is reproducible. Finally, you will submit a **brief** paragraph on each of your models. This paragraph should discuss how your features were chosen as well as the performance of your model.

Models

For the **logistic regression model** component, you will be predicting **qualAboveAverage (excluding 5) in the OverallQual column**. Note the predicted column is a binary column with a value of either 0 or 1 (or False / True, whichever one you prefer). You cannot use OverallQual as a feature. This model must be **40%** more accurate than the baseline.

For the **SVM model** component, you will be predicting **HouseStyle**. This model must be **75%** more accurate than the baseline.

For the **decision tree model** component, you will be predicting **Neighborhood**. This model must be **135%** more accurate than the baseline (although this may seem daunting, neighborhood has a very low baseline).

What to Submit:

A jupyter notebook containing

- any code you used to decide upon features (optional)
- for logistic regression, SVM, and decision tree:
    - model code
    - 5 different train/ test splits (you need to include 5 different **random_state** values you used) and their accuracy compared to the baseline
    - one paragraph on your model (can be in a pdf)

The CSV you created in the previous project and imported for this one. If you don't submit one we will assume you just used the original train.csv.

Criteria	Points
Logistic Regression	
Correct Model - Model is classifying with features	5
Correct Formula - Split baseline and prediction accuracy calculation	5
Accuracy: - Split is 40% more accurate (for all 5 train_test_split's) - (40% = 10/10, 36% ~ 9/10, 32% ~ 8/10, and so on)	10
Paragraph explanation	5
SVM Classification	
Correct Model - Model is classifying with features	5
Correct Formula - Split baseline and prediction accuracy calculation	5
Accuracy: - Split is 75% more accurate (for all 5 train_test_split's) - (75% = 10/10, 67.5% ~ 9/10, 60% ~ 8/10, and so on)	10
Paragraph explanation	5
Decision Tree Classification	
Correct Model - Model is classifying with features	5
Correct Formula - Split baseline and prediction accuracy calculation	5
Accuracy: - Split is 135% more accurate (for all 5 train_test_split's) - (135% = 10/10, 121.5% ~ 9/10, 108% ~ 8/10, and so on)	10
Paragraph explanation	5
Total	**75**