CS 5830-Project 7

Logistic Regression and SVM

Pouyan Saeidian | Megh kc

Presentation link: here
GitHub link: here

Analysis 1 (Megh KC)

Introduction

The wine quality dataset analysis majorly focuses on wine quality concerning its parameters. The quality of wine can vary if there are even simple changes in the ingredients and their quantities. The consumers might be giggly deviated because of the wine quality. Hence, this analysis gives the significance of the chemicals and ingredient qualities to maintain quality. The major stakeholders would be wine consumers. The company can use this analysis to improve its wine products with the best quality, which also affects market penetration.

Dataset

This dataset is related to red variants of the Portuguese "Vinho Verde" wine. The dataset was accessed from kaggle.com. It has been said that the dataset contains limited input & output variables because of privacy issues. Therefore, the dataset was analyzed and classified by using available attributes. The attributes containing the acidity parameters, PH, sulfates, sugar, chlorides & alcohol content, and quality of wines were assigned in numerical integer values. Data has the shape of 1599*12.

Analysis Techniques

The data extracted from Kaggle.com was analyzed using a python environment. The 12 different columns were available, and we added a new index column. The qualities of wines were in integer values ranging from 3 to 8. We modified this quality scale to bad, medium, and good quality scales by using current values (bad- 3,4, medium- 5,6, and good -7& 8) based on the dataset description by some users. The numerical values for modified wine qualities were 0,1 & 2, respectively. The heatmap plot was created to see the correlation among the data attributes. Also, a kde plot for the different 6 attributes has been created to find whether the used attributes would have significance in our analysis. In addition, several scatter plots were created for different qualities. To predict the different qualities, we did one-hot encoding for the quality column where bad quality was a reference. We make training and test dataset for the wine quality prediction with 80% dataset as the training dataset. The logistic regression for predicting medium-quality wine through the alcohol content & residual sugar was done, which gave us a decent f-score of [0.039, 0.91].

Similarly, the high-quality wine classification was predicted using the fixed acid & volatile acid contents with a class weight of 0.15 and produced a score of [0.82, 0.39]. We saw that the decision boundary on high-quality wine classification by using the fixed acid and volatile acid was significantly impacted by class weight. Therefore, we used class weight 0.15, which gave us a nice boundary. We used SVM for the same classification above (high-quality wine with fixed & volatile acid content) with linear and polynomial kernel and found a reasonable f-score [0.84, 0.42] & [0.90, 0.16]. But, the class weight here also had significant fluctuations. We used class weight 1:4 for 0 & 1 values in this quality class. The radial bias function for the same analysis with the same weight gave us a different value than the degree 5 SVM polynomial classification with an f-score [0.87, 0.33].

Result

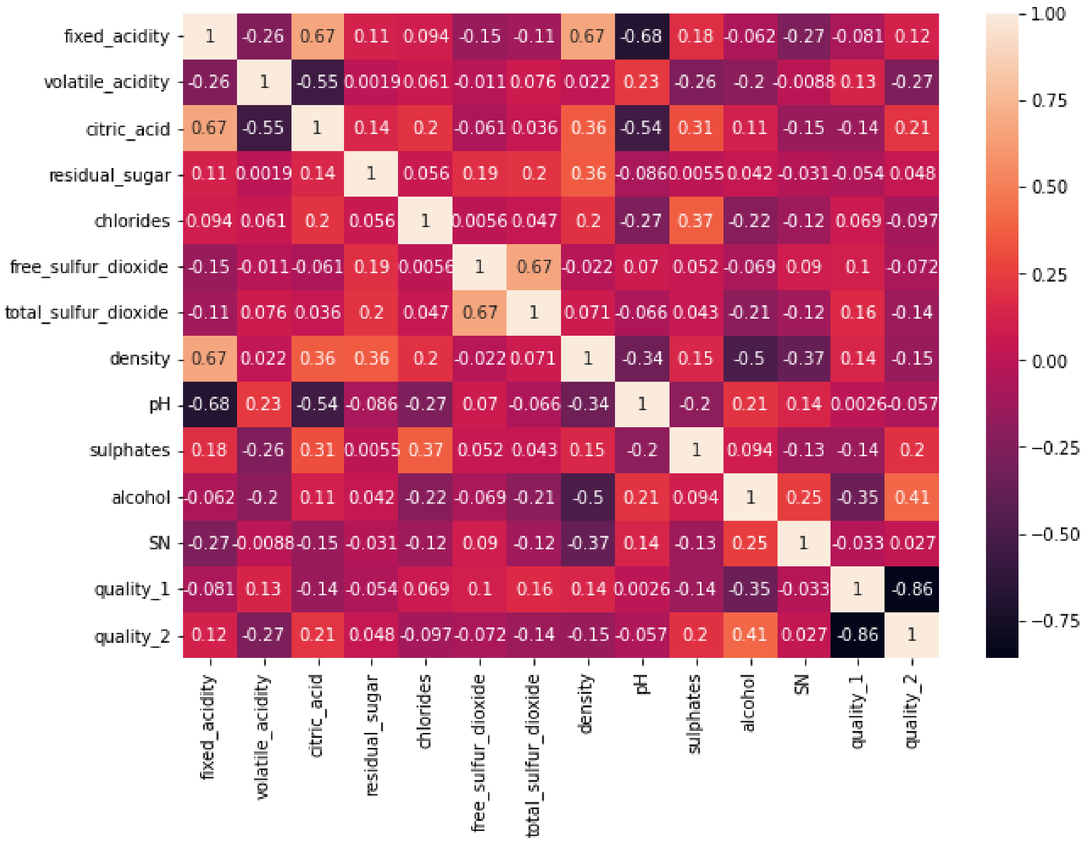
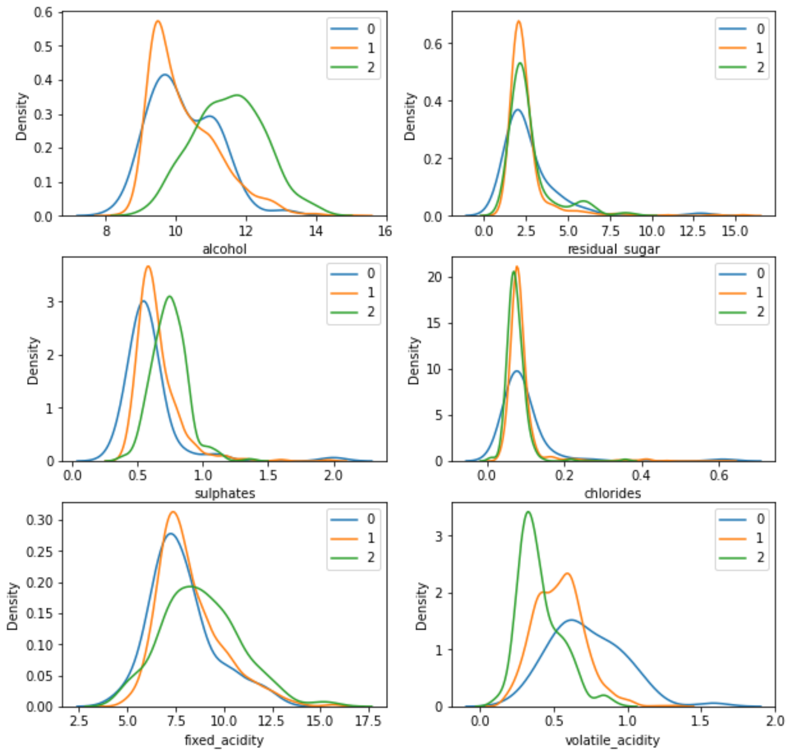
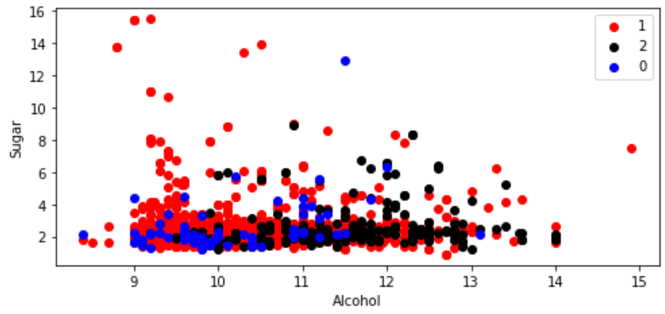
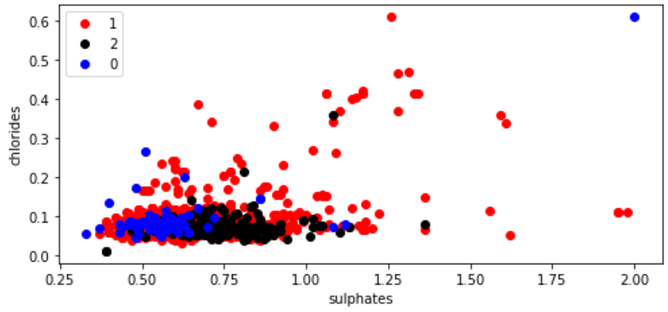
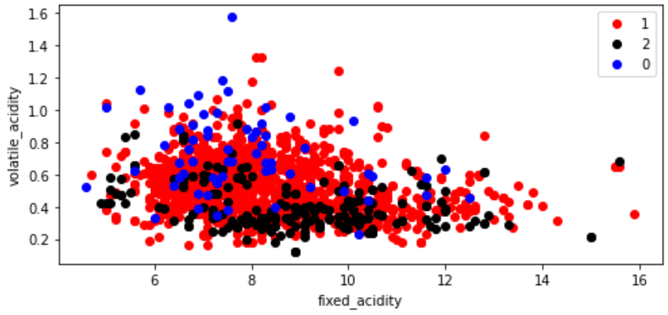


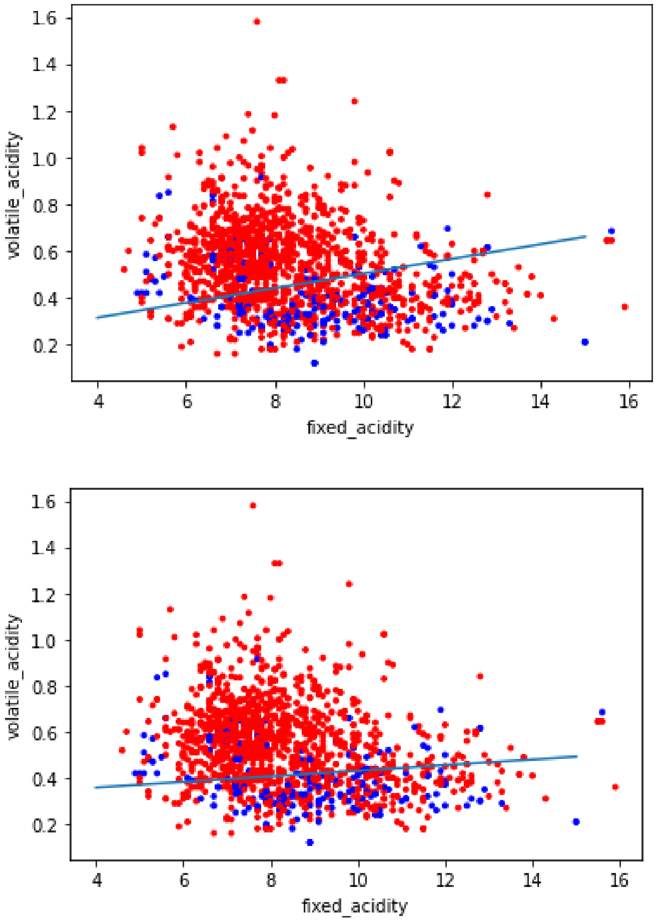
Fig: correlation heatmap


Density
alcohol
residual_sugar
sulphates
chlorides
fixed_acidity
volatile_acidity
0
1
2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
8
10
12
14
16
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20
0.05
0.10
0.15
0.20
0.25
0.30
1.5
2.0


Sugar
Alcohol
1
2
0
2
4
6
8
10
12
14
16
9
10
11
12
13
14
15


chlorides
sulphates
1
2
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00


volatile_acidity
fixed_acidity
1
2
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
6
8
10
12
14
16



a) Logistic regression fit line, b) linear SVM

Results Table for high-quality wine for fixed and volatile acid content

SN	Analysis	f-score	remarks
1	Logistic reg	[0.82,0.39]	weight 0.15
2	Linear SVM	[0.84,0.42]	(0:1,1:4)
3	Polynomial SVM	[0.90,0.16]	(0:1,1:4), degree 5
4	RBF	[0.87,0.33]	(0:1,1:4)

Technical:

We did one hot encoding for quality classes to create a logistic regression for each class. An SVM was suitable for this dataset since we did a binary classification. The polynomial kernel worked the best since a line could separate the data. We did not find any significant difference between the different models. However, polynomial SVM works best for predicting negative values. For linear SVM, we tried different weights between 1 to 7; the difference was significant. The weight of 1:4 worked well. The decision boundary hyperplane for the prediction of high-quality wine using acid intensity is shown in the figures in the Result section. The dataset is relatively classifiable once good sets of attributes are picked.

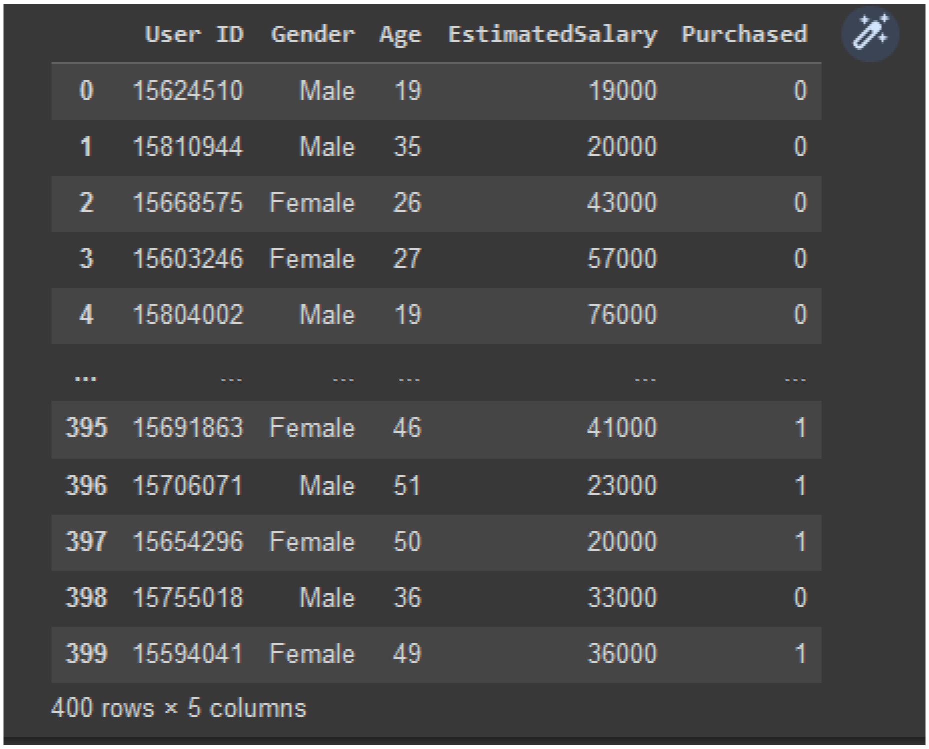
Analysis 2

Introduction

The analysis in this project aims to identify the variables that influence consumer purchasing decisions on social media. Understanding these motivations is crucial for understanding consumer behavior on social media and improving marketing strategies.

Dataset

The Social_Network_Ads.csv dataset on kaggle.com is a dataset with information on social network users and whether or not they purchased a particular product advertised on the social media platform. The dataset includes information about each user's age, estimated salary, gender, and the outcome of the advertising campaign (whether they made a purchase or not).



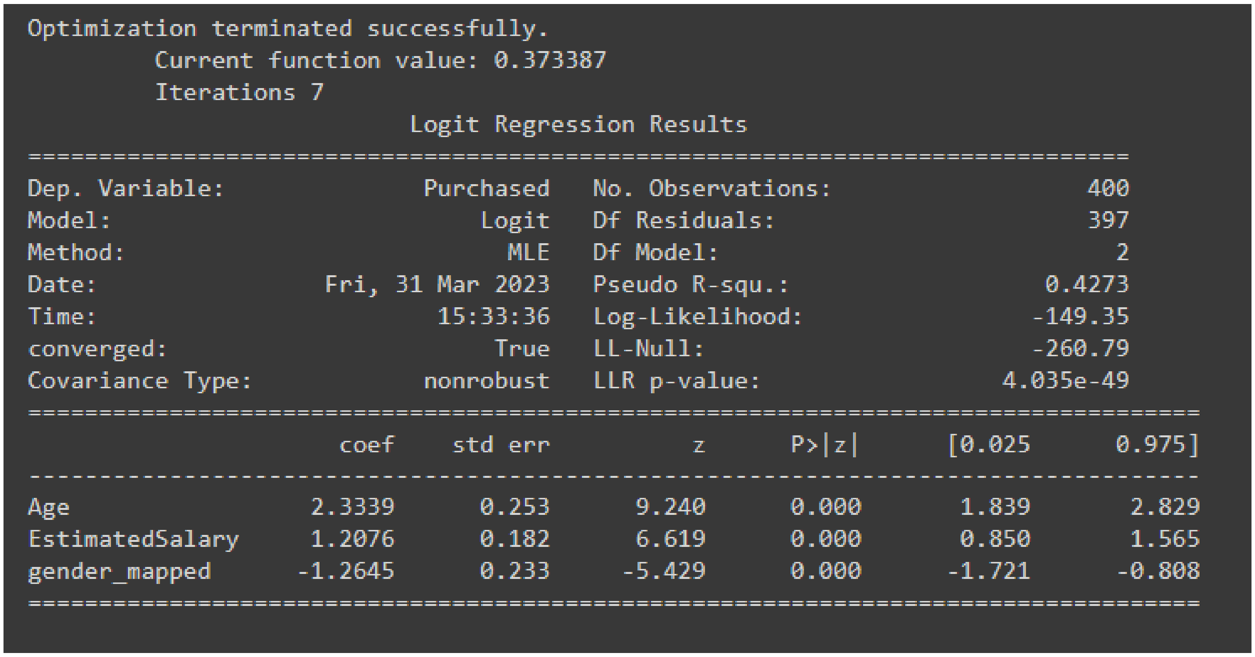
	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0
...	...	...	...	...	...
395	15691863	Female	46	41000	1
396	15706071	Male	51	23000	1
397	15654296	Female	50	20000	1
398	15755018	Male	36	33000	0
399	15594041	Female	49	36000	1

400 rows × 5 columns

Analysis techniques and Results

We are trying to model which customers will purchase a product after seeing an advertisement. The target variable is a binary purchase, and the independent variables are income, age, and gender (predictive of target variables).

We have used logistic regression and support vector machine(SVM) with different kernel types (Linear SVM, Polynomial SVM, and RBF) for the classification purchase phase. The result of logistic regression is shown below.



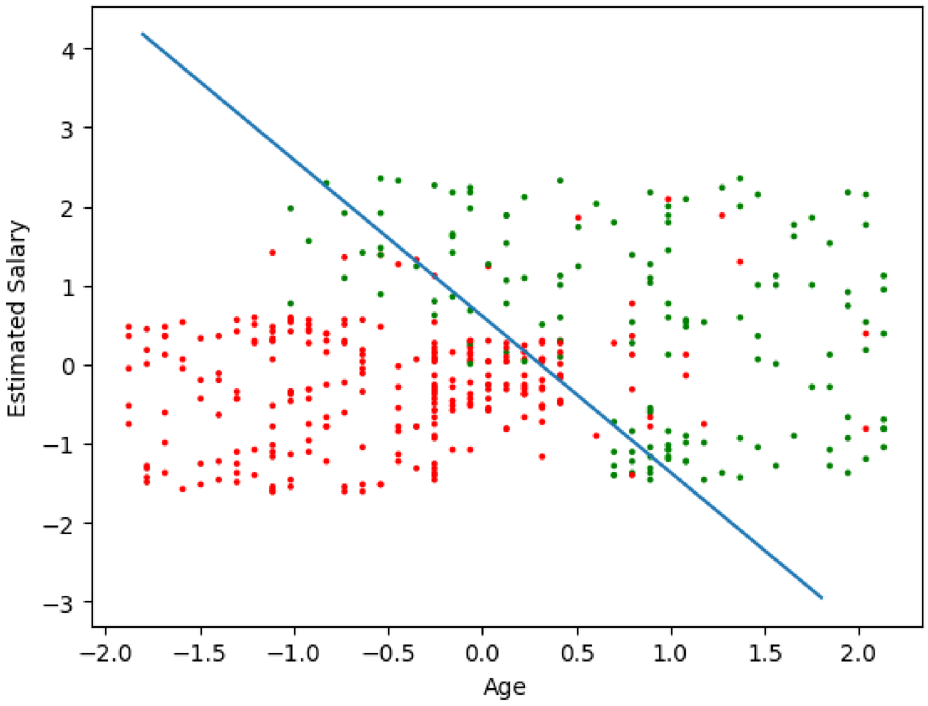
```
Optimization terminated successfully.
         Current function value: 0.373387
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:              Purchased   No. Observations:                  400
Model:                          Logit   Df Residuals:                      397
Method:                           MLE   Df Model:                            2
Date:                Fri, 31 Mar 2023   Pseudo R-squ.:                  0.4273
Time:                        15:33:36   Log-Likelihood:                -149.35
converged:                       True   LL-Null:                       -260.79
Covariance Type:            nonrobust   LLR p-value:                 4.035e-49
===================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
Age                 2.3339      0.253      9.240      0.000       1.839       2.829
EstimatedSalary     1.2076      0.182      6.619      0.000       0.850       1.565
gender_mapped      -1.2645      0.233     -5.429      0.000      -1.721      -0.808
===================================================================================
```

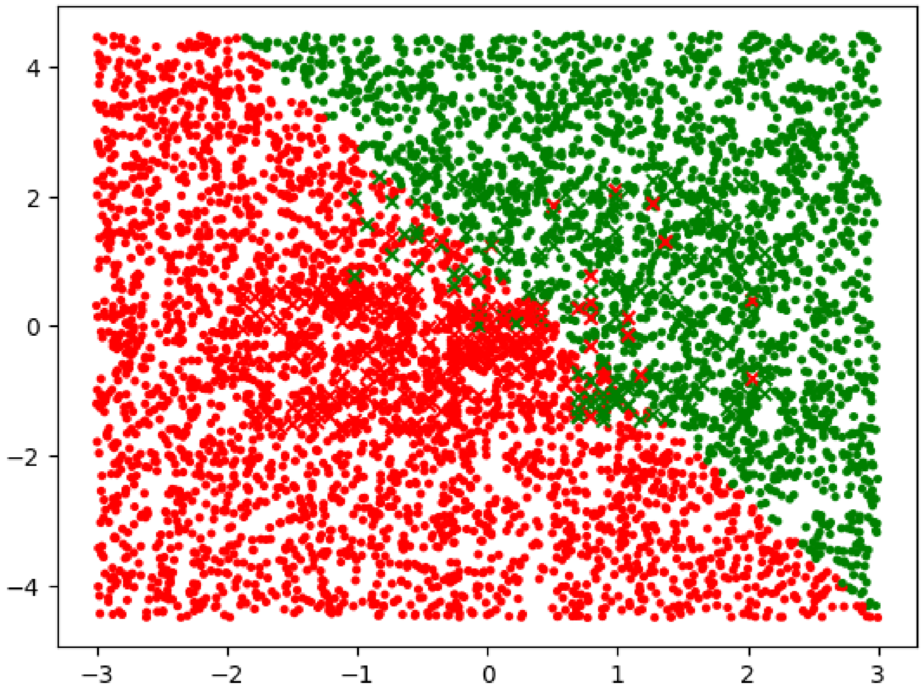
Can logistic regression or a linear SVM predict well?

We trained and tested a dataset using logistic regression and linear support vector machines on a 25% test set. The logistic regression classifier achieved an accuracy of 0.89 on the test set, while the linear support vector classifier achieved an accuracy of 0.87. Our results show that the logistic regression classifier outperformed the linear SVM regarding prediction accuracy.

What do plots of selected pairs of variables look like? Where is the decision boundary in those plots?

We can observe the decision boundary that separates the two variables in the logistic regression, but it misclassifies some blue data points. The decision boundary in the SVM is similar to that of the logistic regression model. To better understand the data and boundary line, we drew similar graphs for the SVM with a Polynomial and Radial bias function and populated them with random points.

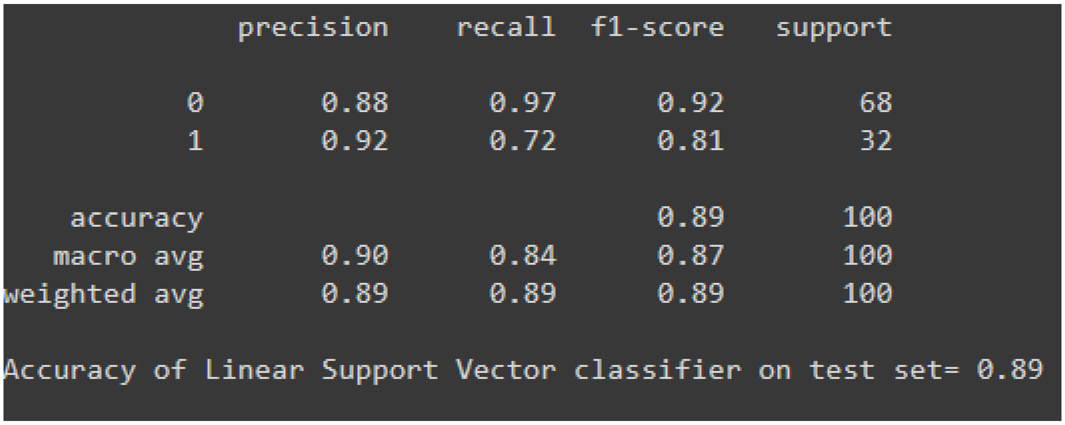




How generalizable are the different models on your data? How does the bias-variance tradeoff affect which model you might choose?

Our models have demonstrated strong generalizability, as evidenced by their high f-score statistics on our test datasets. We have selected the model with the highest accuracy and the lowest time for use.

Regarding the Bias Variance Tradeoff, our model is relatively simple, containing only three variables and two classes, and therefore has high bias and low variance. However, if our model had many parameters, it could have had low bias and high variance.



```
              precision    recall  f1-score   support

           0       0.88      0.97      0.92        68
           1       0.92      0.72      0.81        32

    accuracy                           0.89       100
   macro avg       0.90      0.84      0.87       100
weighted avg       0.89      0.89      0.89       100

Accuracy of Linear Support Vector classifier on test set= 0.89
```

Is there a difference between the polynomial and RBF SVMs?

On the test set, the Polynomial Support Vector classifier achieved an accuracy of 0.85, whereas the RBF Support Vector classifier achieved an accuracy of 0.90, the highest achieved so far.

What effect does changing the class_weight in an SVM have on your data? How might this be important for this data?

The number of samples that were purchased and not purchased. I have considered weight once at 0.4 and the other times at 0.5.

Is there a difference in runtime performance?

We can observe that the runtime varies for each analysis, with Logistic Regression and Radial Basis Function taking the least time. Conversely, SVM (Linear) took the longest by far.

Logistic regression and linear SVC use one-vs-rest (OVR) for multi-class classification. SVC uses one-vs-one (OVO). Where n is the number of classes, OVR learns n models, whereas OVO learns n(n-1)/2 (n choose 2) models. What effect does this have on performance?

The OVR approach is expected to be faster than OVO because it learns fewer models.

In contrast, OVO has to learn n(n-1)/2 (n choose 2) models, which is a more significant number than n itself.

Technical

We have converted category gender to numeric values 0 and 1 and then ran the logistic regression and evaluated its performance. After that, we implemented SVM with different kernel functions and compared the performance of each method.