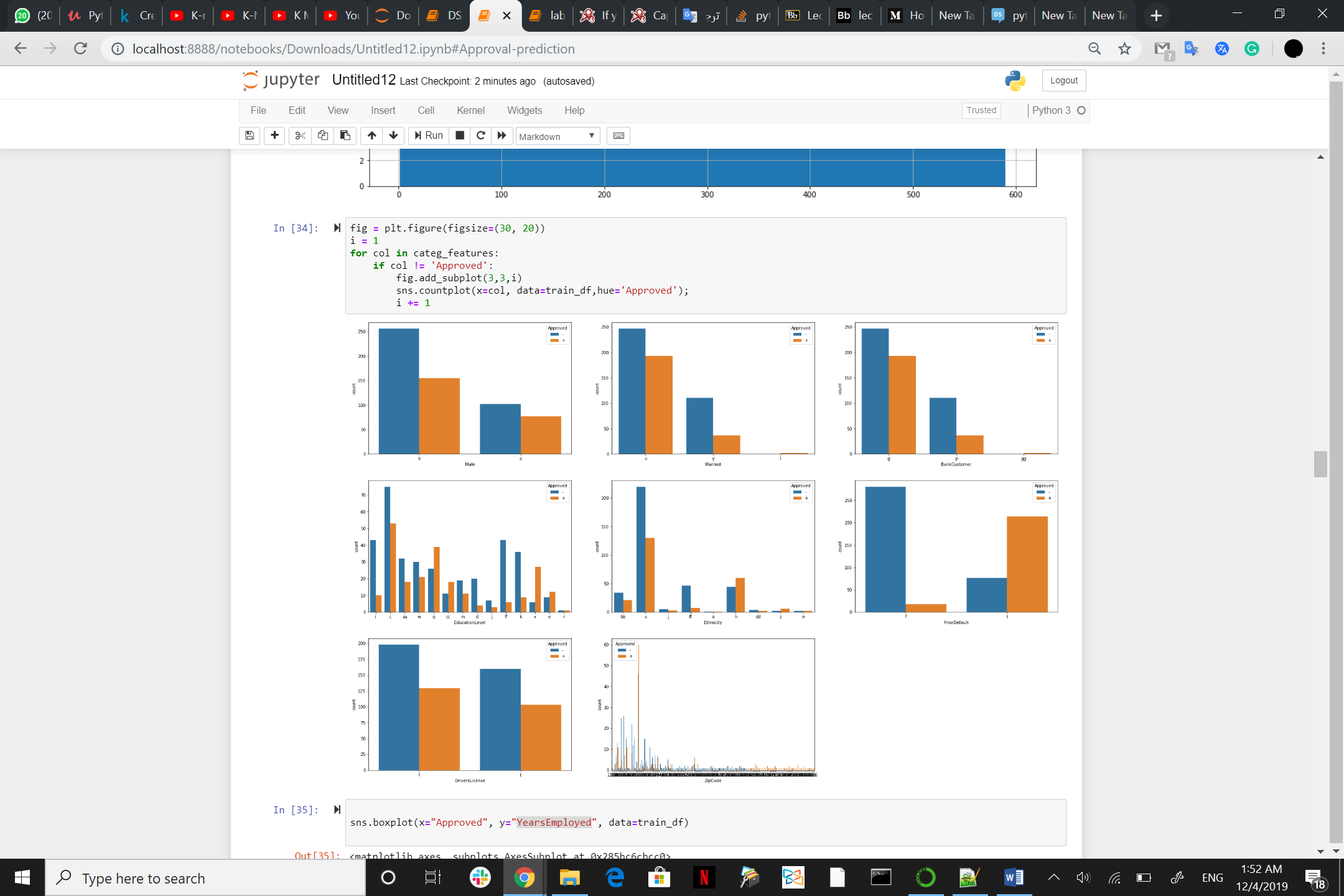
|  |  |
| --- | --- |
|  |  |

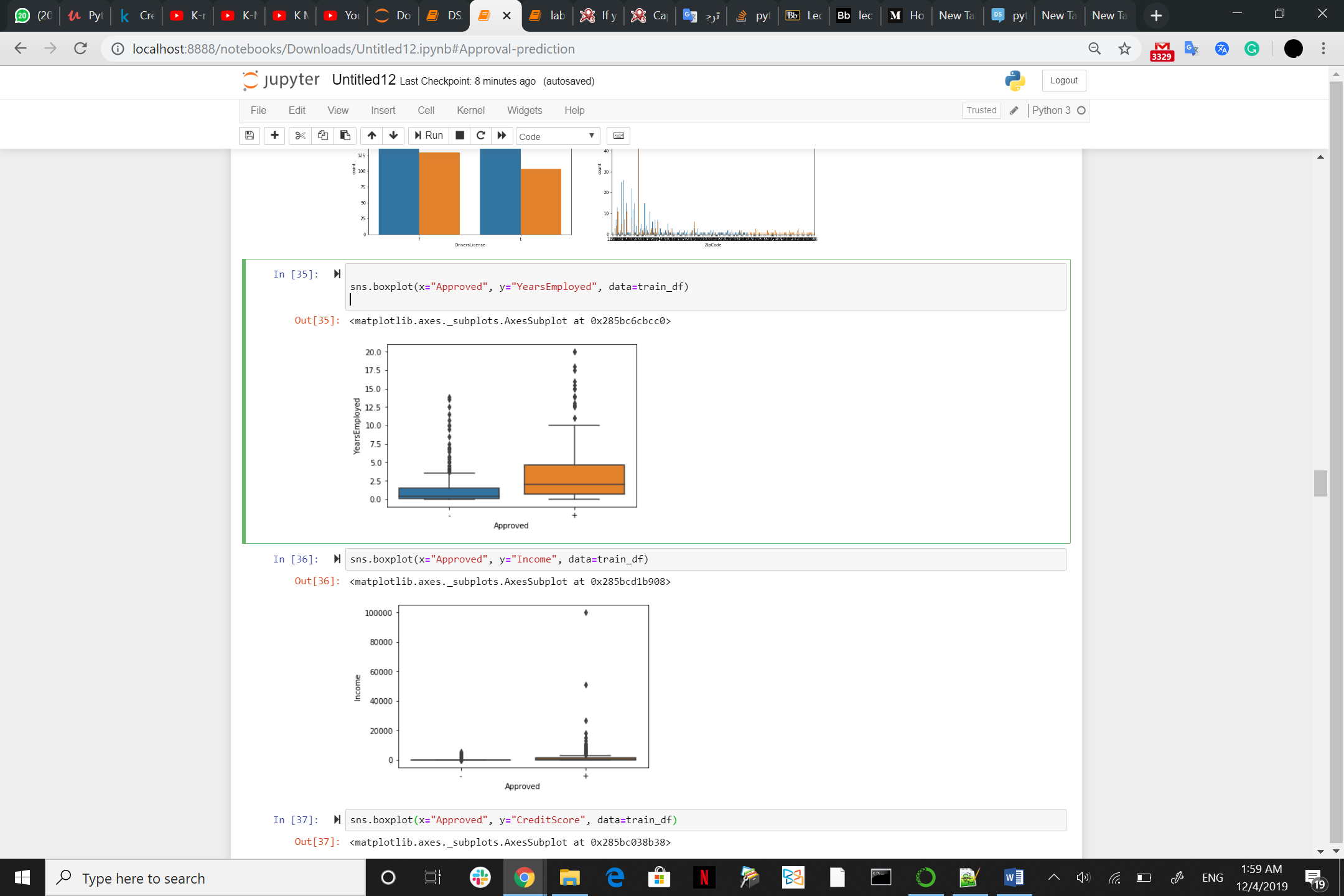
**Assignment 3: Modelling**

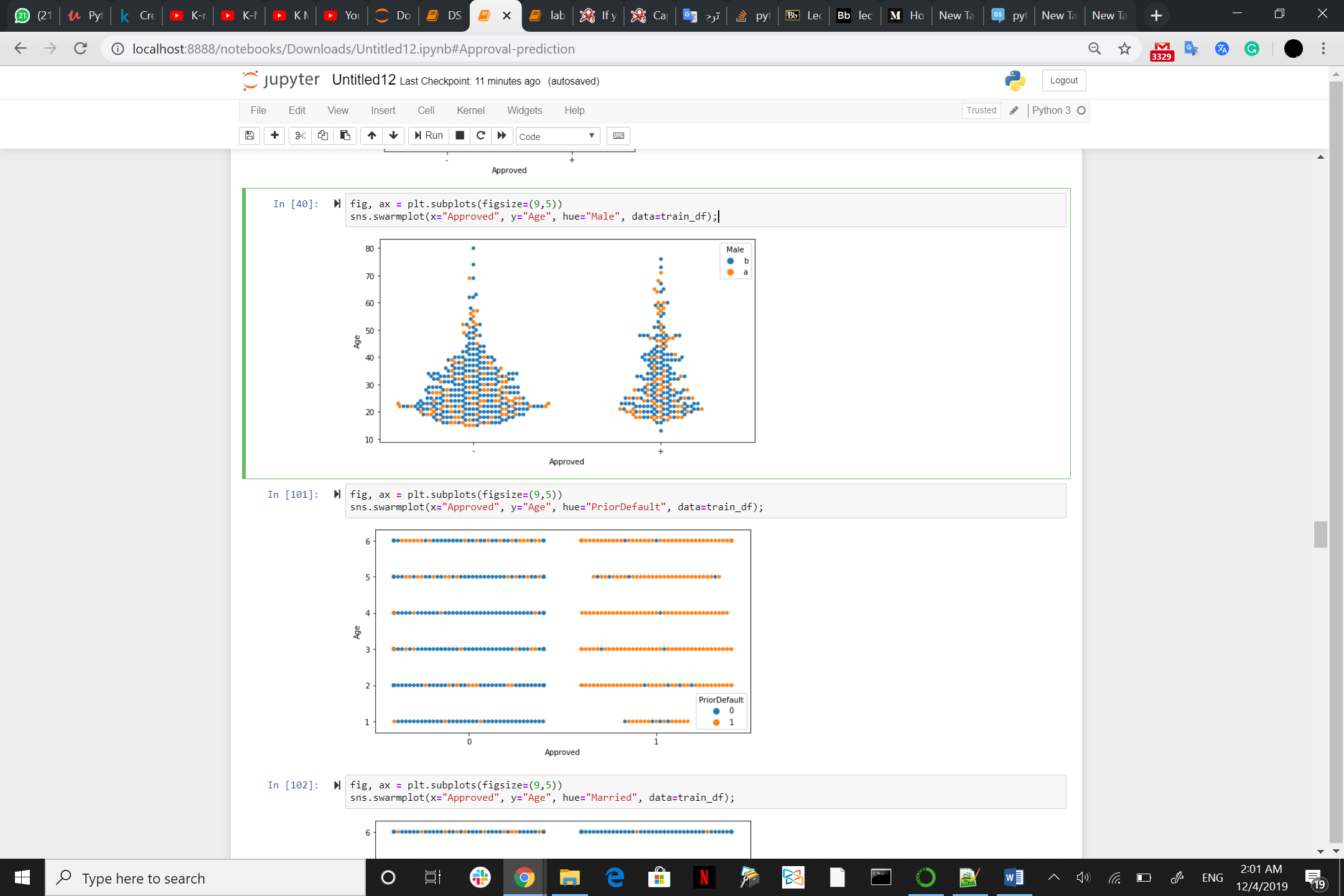
**EAD and Data preparation**



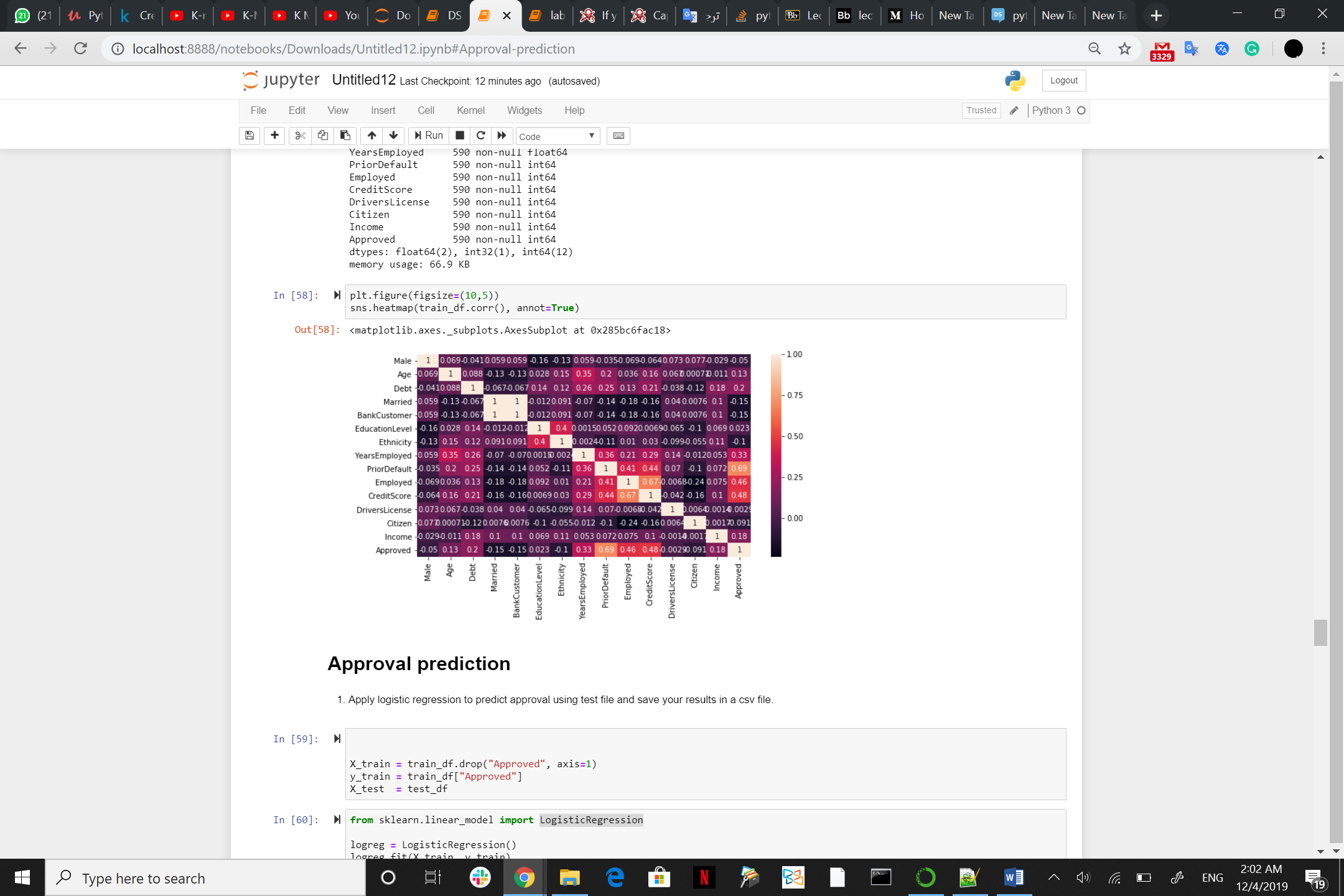
Plotting everything against Approved

As you see the ‘PriorDefault’ column matter.



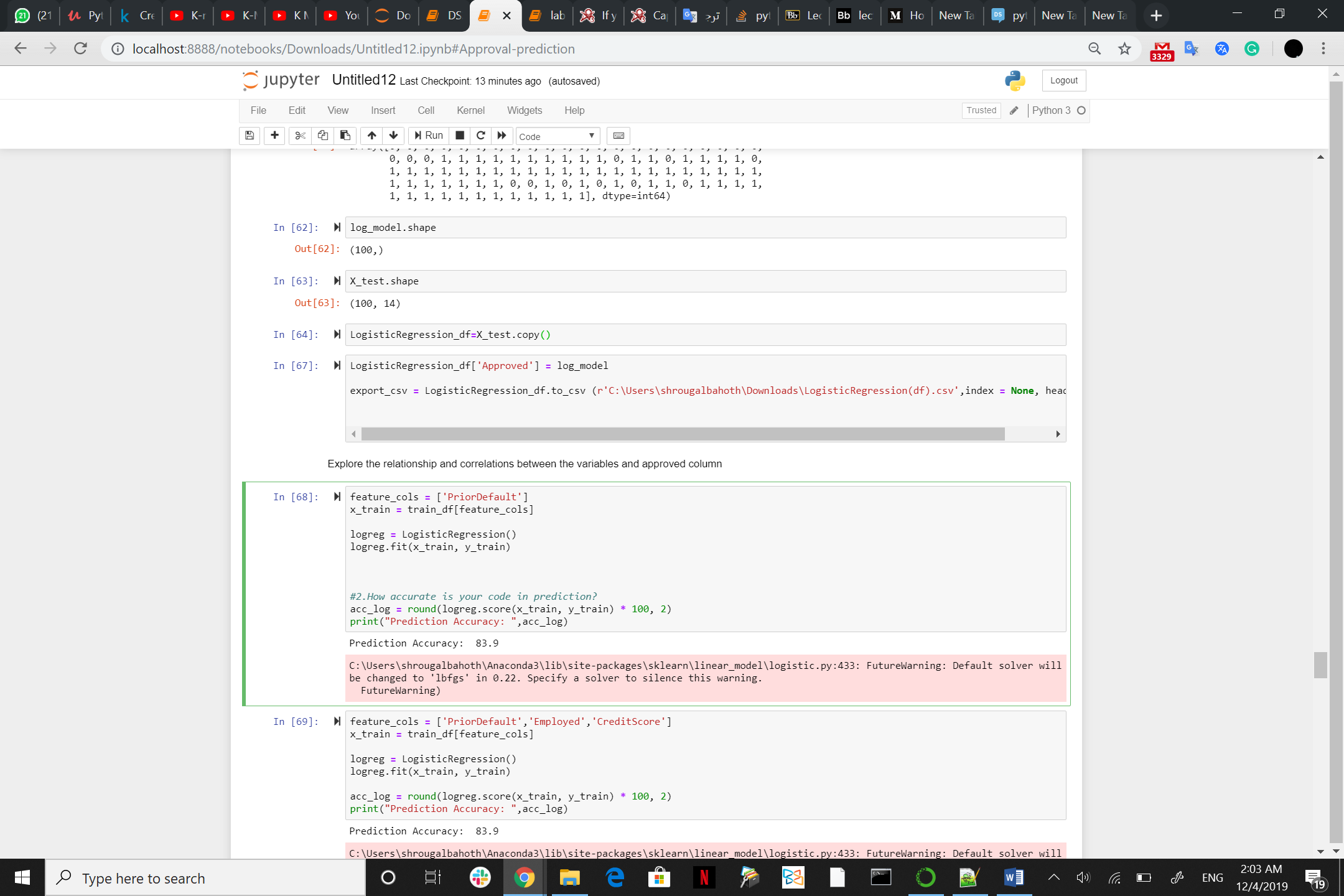


Here also you can see how ‘PriorDefault’ column matter



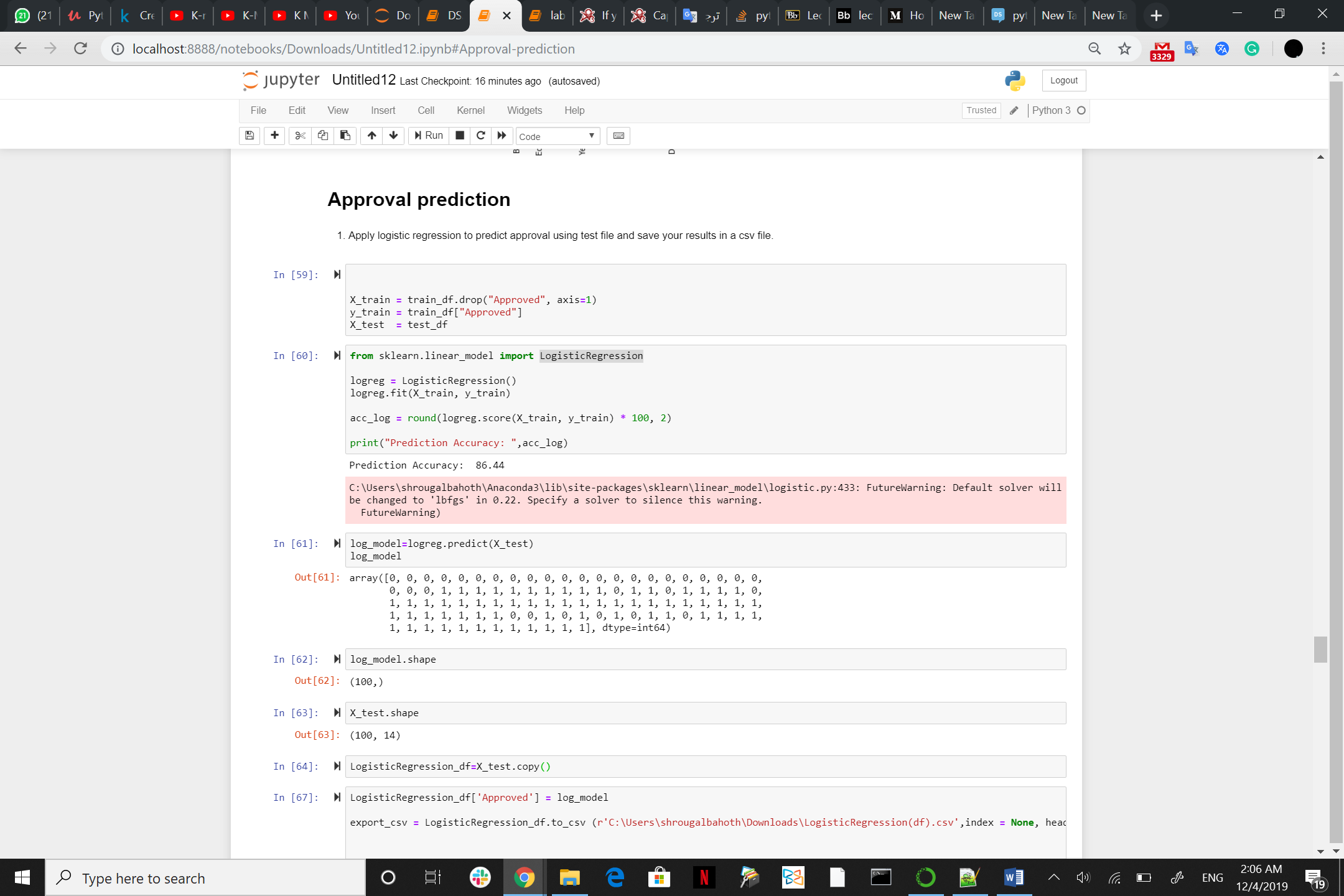
there is a 0.69 correlation between Approved column and PriorDefault column we can consider at as a strong relation .

also, there is a 0.46 correlation between Approved column and Employed column and a 0.48 correlation between Approved column and CreditScore column

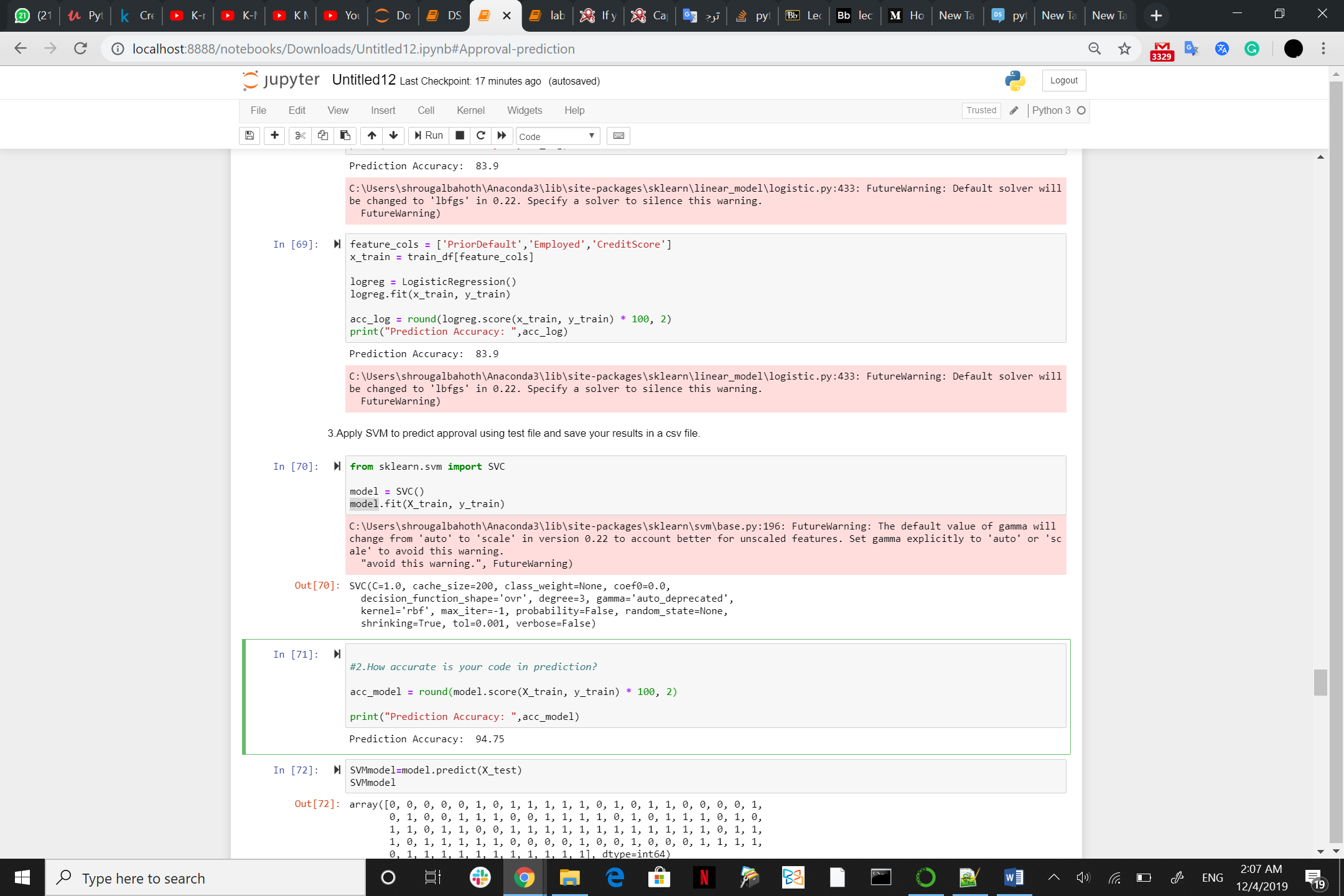


PriorDefault is highly significant in predicting approval

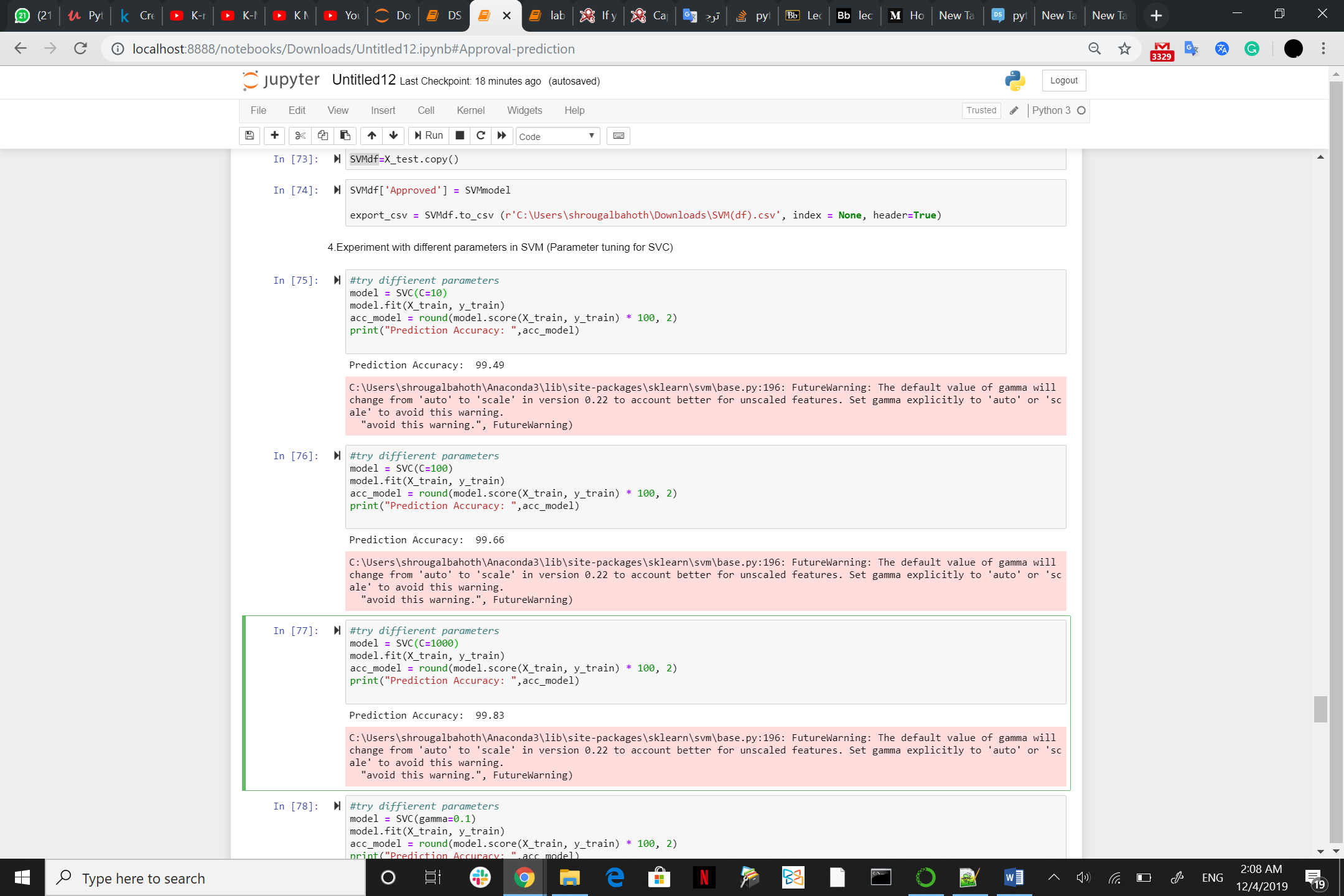
**Approval prediction**

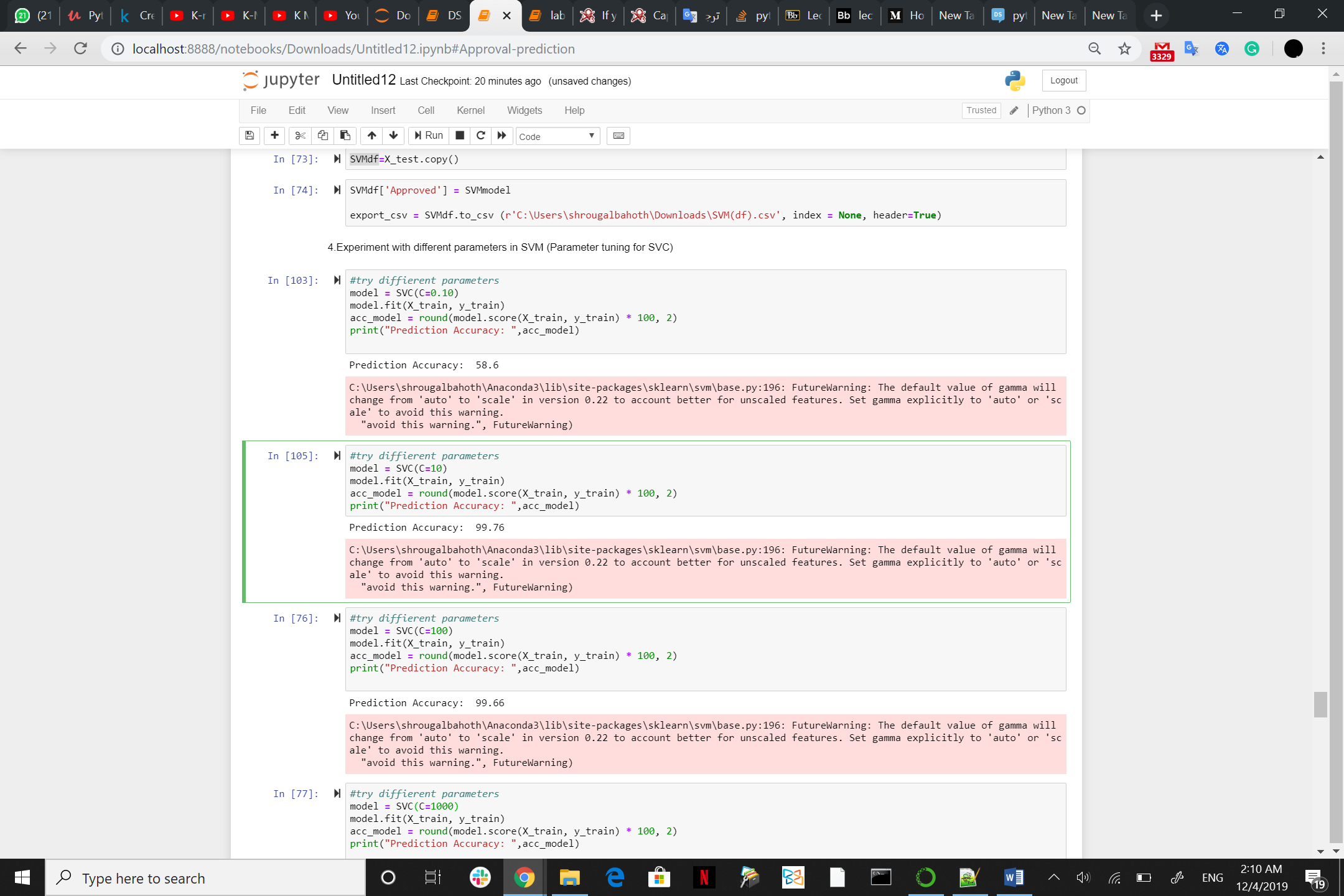


Prediction Accuracy applying logistic regression is 86.44%



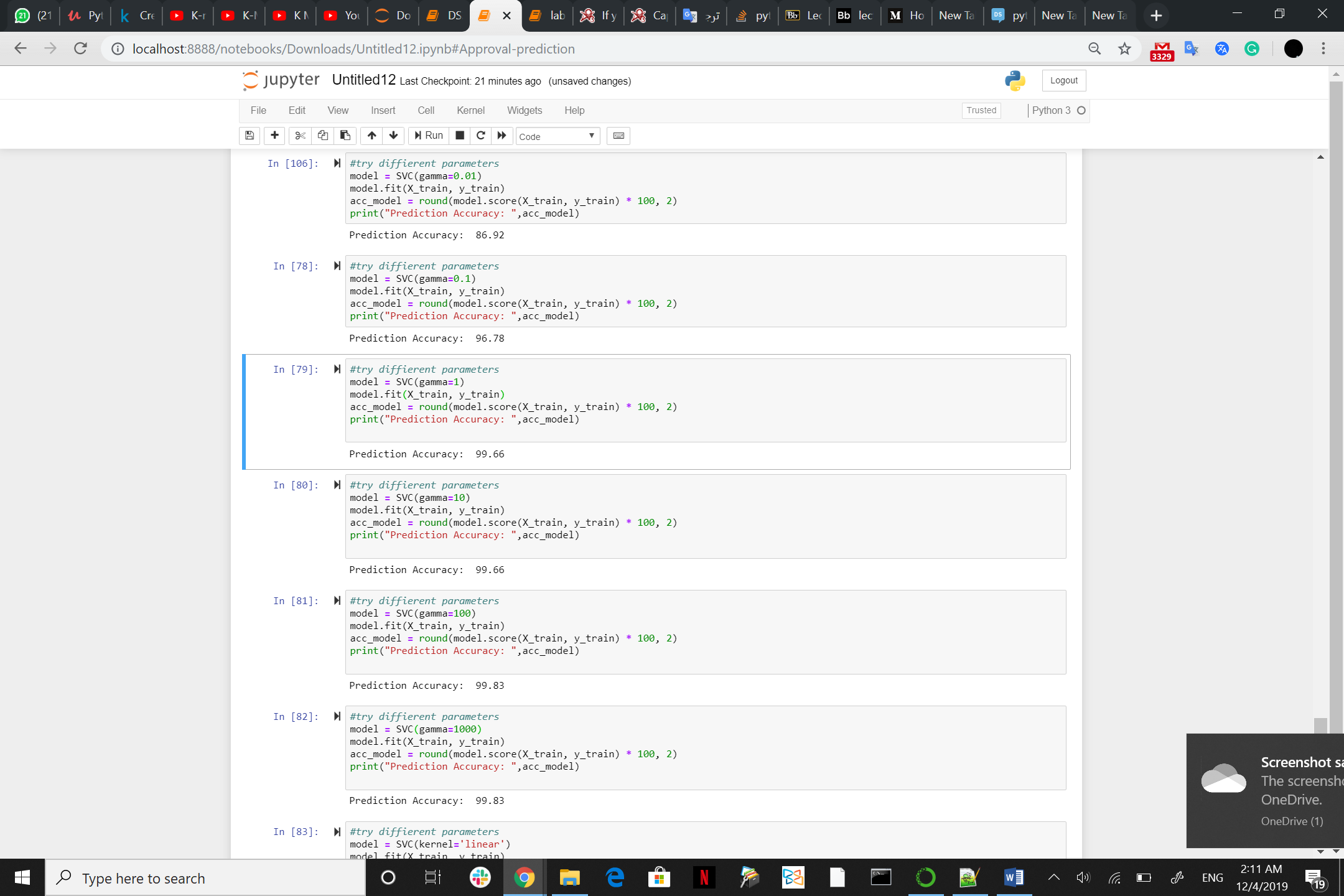
Prediction Accuracy applying SVM is 94.75%





C is the penalty parameter of the error term. It controls the trade-off between smooth decision boundary and classifying the training points correctly.

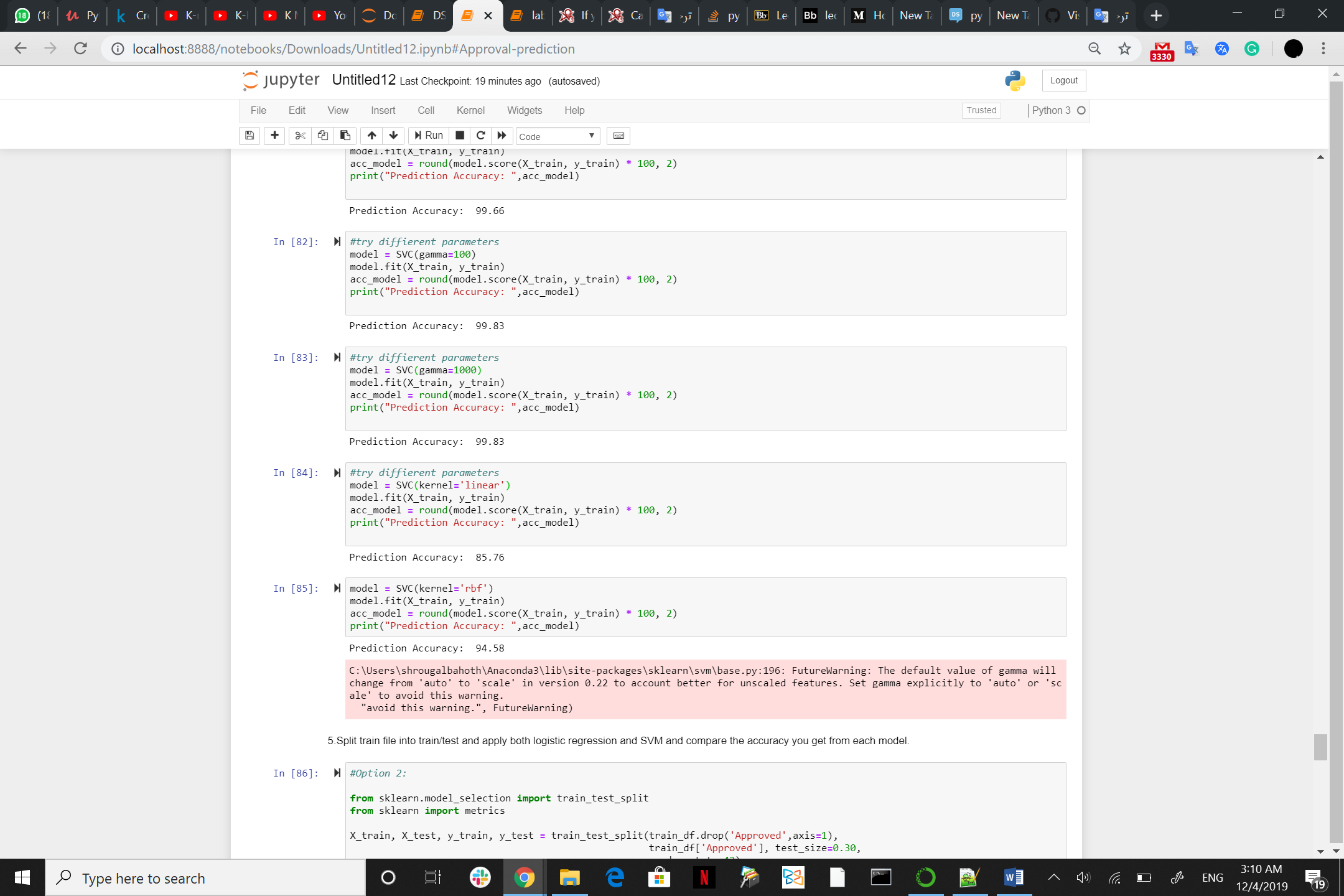
When I increase C parameter value prediction accuracy is increasing as well



gamma is used for non-linear classification problems. It tells the influence of one training example on the others. Smaller gamma means high influence and vice-versa

 the gamma parameter defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. The gamma parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors.

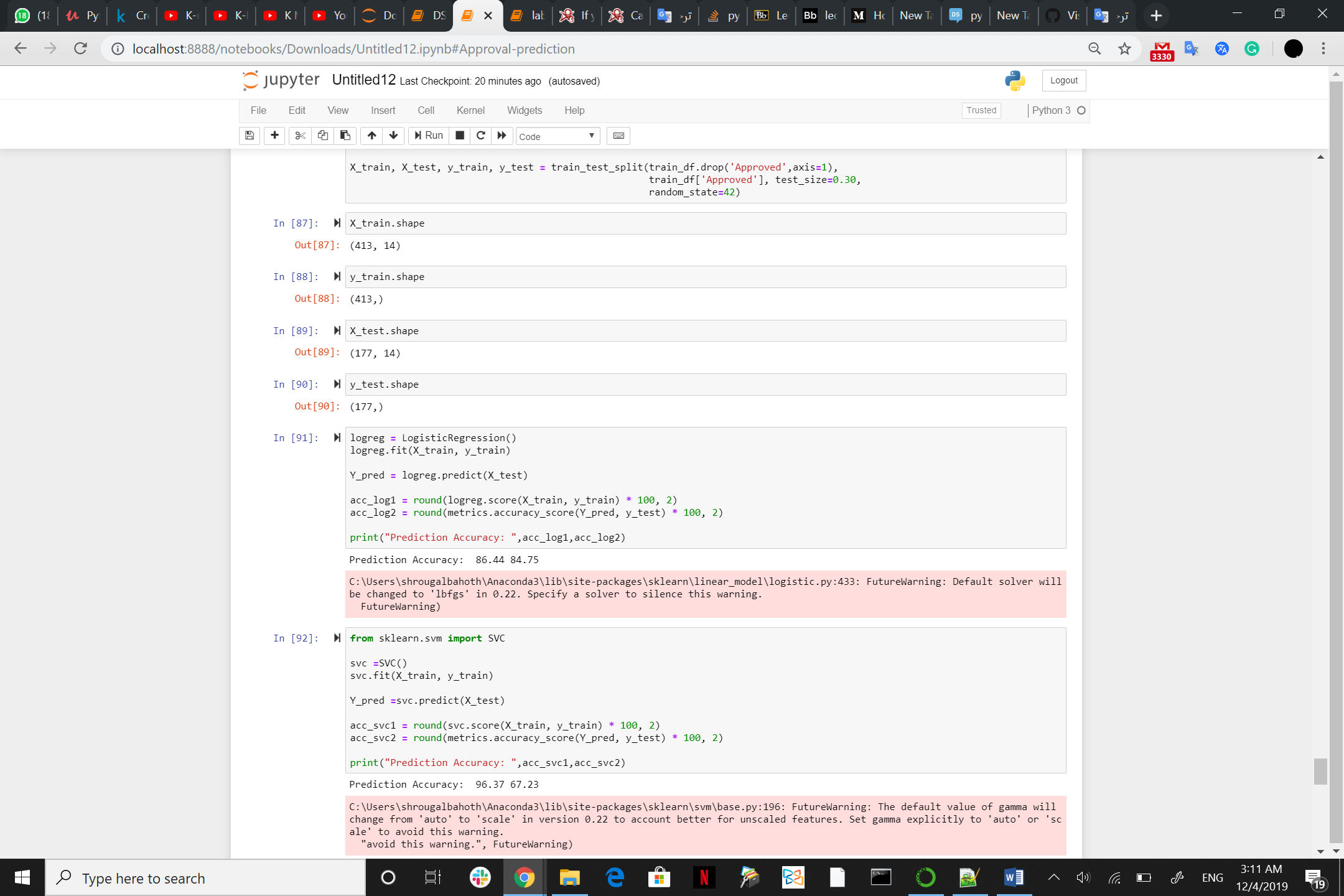
When I increase gamma parameter value prediction accuracy is increasing as well.



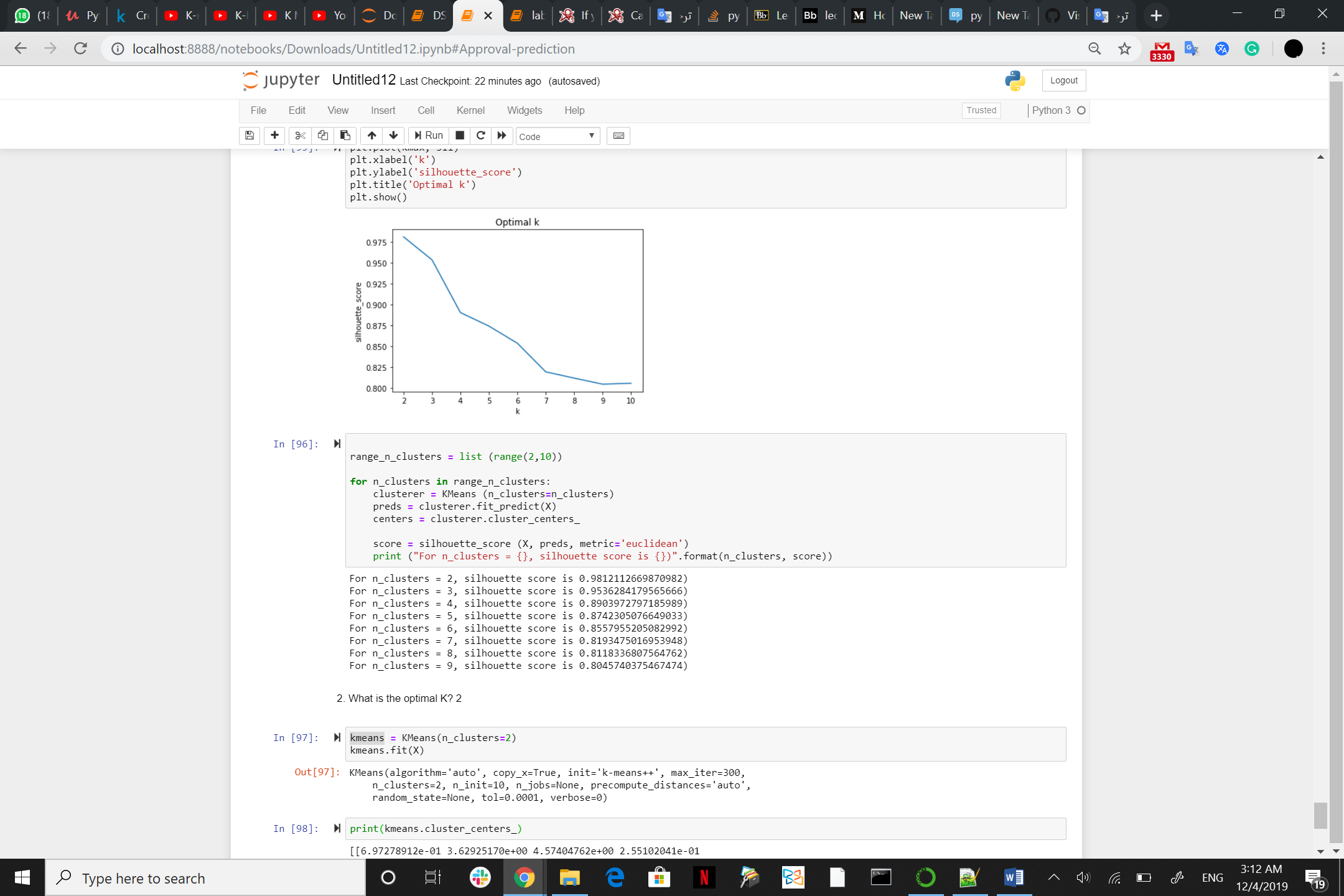
Linear Kernel is used when the data is Linearly separable, that is, it can be separated using a single Line.

As you see RBF Kernel in this case provided higher prediction accuracy

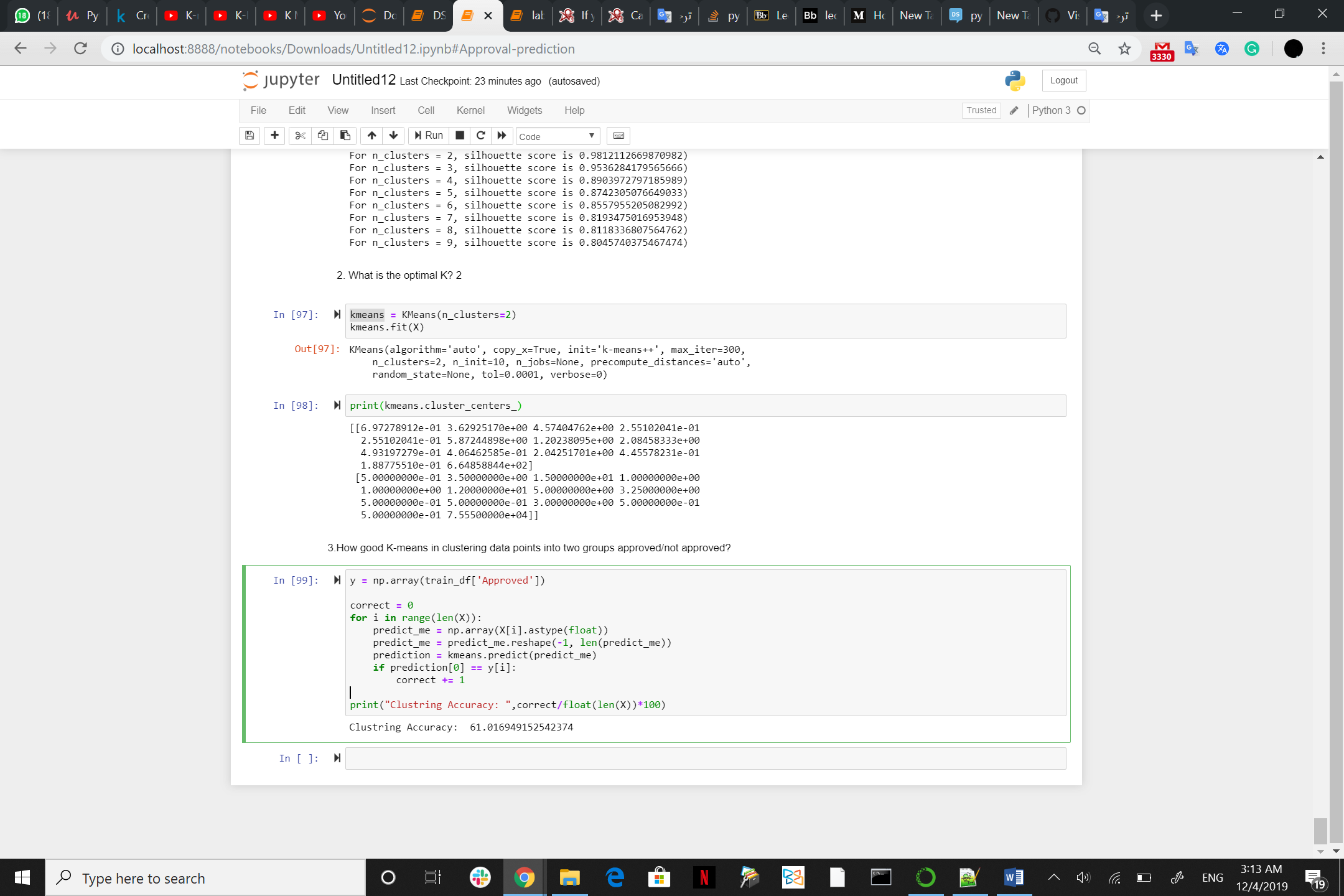
# Split train file into train/test and apply both *logistic regression* and *SVM* and compare the accuracy from each model.



**Clustering**



I applied K-means clustering to group data points, and I used silhouette coefficient to find the optimal number of K for the dataset.The optimal K as you see is 2.



The accuracy of K-means in clustering data points into two groups approved/not approved is 61.01

The biggest disadvantage is that K-Means requires you to pre-specify the number of clusters (k).and it is sensitive to outliers and different results can occur if you change the ordering of the data.