INTL 550: Hw3 - Emre Uzel

I have started my analysis reading and looking at features. I checked if there are illogical or null values in the dataset. There were no null values but there were some illogical ones. When I've found illogical ones, I converted them first 'np.nan'. Then, I change them to median of the feature they belong using Imputer. After that, I encoded the variables that I need to encode, using one-hot encoding.

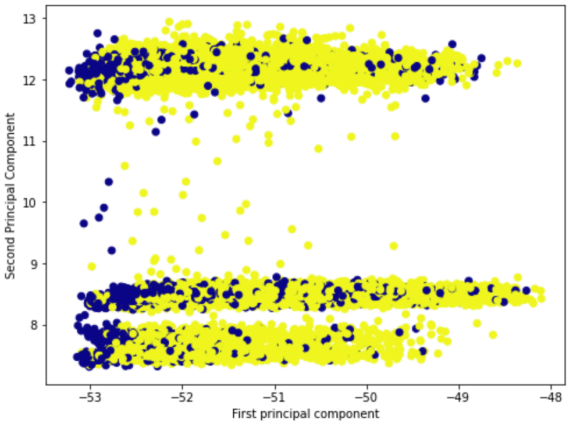
My strategy for ML is going from easiest to hardest. I first implement algorithms on raw data. Then implement same algorithms using feature selection and normalized version of data to see if there is a difference. I recorded every result and I will show them in a table below. My main comparison is AUC score. I belive that using AUC score is much more logical because 80% of data is True(1) so it's better to identify whether the algorithm is working well. Normalization didn't have effect on Tree models (RandomForest, GradientBoosting) however it has improved performance of SVM.

Choosing few important features, improved performance of tree models.

I have implemented 4 algorithms; RandomForestClassifier, GradientBoostingClassifier, SVM, LogisticRegression. RandomForest performed best and that's why I implemented tuning on RandomForest. My tuning results performed worse in terms of AUC but it has increased performance in terms of accuracy. That's why according to business objective we can use either two model.

	SVC	Random Forest	Gradient Boosting	Logistic Regression
Raw_Implementation	0,58	0,642	0,664	0,631
Raw_Feature_Selected	0,55	**0,671**	0,639	0,59
Standardized	0,57	0,641	**0,663**	**0,632**
Standardized_Feature_Selected	**0,596**	0,669	0,639	0,59
Tuned		0,645		

I have implemented PCA as well. But the results are hard to distinguish and I believe that it is not identifying 1 and 0 clearly as you can see in the figure below. That's why I didn't implement algorithms after I implement PCA.



Lastly, I've worked on Jupyter notebook. That's why I am both attaching .ipynb version and .py version as well. If you could open it in Jupyter or I python, it would make more sense and you can see the steps I took much clearly.