Advanced Machine Learning

Module Code: MOD006566

Academic Year: 2022-2023

SID: 2050132

Trimester: 1

Table Of Contents

Background and Description

This assignment is based on the ‘Spaceship Titanic’ kaggle challenge whose main objective is to obtain a prediction of the number of people who got spirited away into an alternate dimension which occurred as a byproduct of a mishap during one of its voyages from the data which was recovered from the wreckage of the ship’s computer systems. It is almost a dead ringer to its predecessor bearing a similar moniker which sank almost a century earlier, with the exception being that in this case approximately 50% of the individuals aboard the ill-fated vessel were spirited away into an alternate dimension.

Literature Review

Research carried out by the authors of this paper (Vinavatani et al., 2022) involves a detailed study carried out on a variety of applications involving the usage of artificial intelligence to locate missing people which finally concluded in the proposal of a unique mechanism whereby a person's face could be identified with an approximate precision of 90% as compared to the usage of other models to achieve the same but with a much smaller precision.

The research presented in this paper (Shetty et al., 2018) bears a stark resemblance to the description of the problem statement. In this paper, a number of machine learning models have been trained in order to predict the survival rate of the passengers onboard the 'Titanic' namely decision trees, SVM and logistic regression respectively. The best accuracy provided by each of these models culminates in the prediction of the survival rate of passengers.

The research presented by the authors of the paper (Ju et al., 2021) bring to light the various difficulties faced in training models when it comes to big data as traditional models aren't equipped to do so. In order to reduce such difficulties in the near future especially when it comes to classification problems, this paper has put forth the notion of having SVM specific trained models to deal with big data.

This research paper (Kumari & Kr., 2017) puts forth the concept of sockpuppets being one of the many applications of binary classification which itself is defined as a process of classification of a document on the fundamentals of a predefined class. Sockpuppets is an issue which can be resolved with binary classification as any individual can get access to a fake identity and use it for an ulterior or rather malicious motive, which can be fixed via text categorization which is performed in conjunction with binary classification.

The research depicted in this paper (Song et al., 2015) puts forth a concept of a new method which is loosely based on the skeleton of a random forest classifier. Random Forest Classifier has been precisely chosen for implementation in this problem mainly because of its excellent performance and accuracy. This new method is then further utilized in an application known as droplet fingerprint recognition that fashions itself as a method that is used to classify a liquid which after thorough analysis can be used to train the random forest classifier with fingerprint data and finally culminate in the achievement of an extremely high recognition rate.

Data Analysis and Data Pre-processing

Kaggle is the proprietor of the dataset which is associated with the 'Spaceship Titanic' prediction challenge. This dataset is divided into three separate files namely train, test and sample submission. The data present in the training dataset is much more than the percentage of data present in the testing data set owing to the fact that the more the data in the training dataset, the greater is the percentage in the overall performance of the models that are being trained with it. The data that is provided in the training dataset has approximately 8700 entries which mostly include the personal data of the passengers onboard the disaster bound ship.

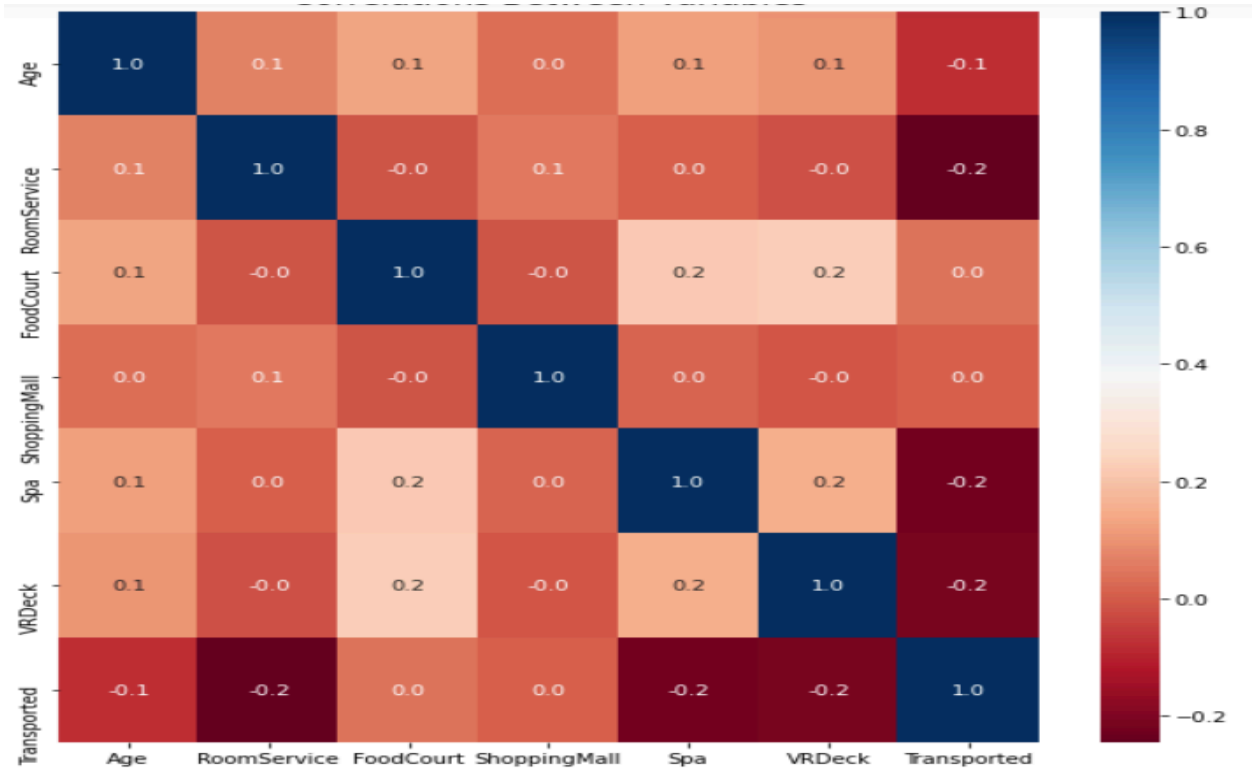
The entire process of machine learning follows a step-by-step approach which starts with the process of acquiring data and ends with the deployment of the final model. The data in regard to the above mentioned problem statement of the 'Spaceship Titanic' is loaded after importing the necessary libraries from 'sklearn' in order to fulfill the necessary criteria for running various models of choice in order to aid in the prediction of the passengers who had been displaced after the ship was struck with an inconsistency which seemed to have safely encompassed itself within the confines of a dust cloud. After acquiring the data, the necessary columns are added in order to make more sense of the data, which is then followed by the implementation of a shape and info method which output the real quantity of elements present within the training dataset along with a distinct structure which consists of the entire structure of the dataset as depicted below.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8693 entries, 0 to 8692
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   HomePlanet    8492 non-null   object
 1   CryoSleep     8476 non-null   object
 2   Destination   8511 non-null   object
 3   Age           8514 non-null   float64
 4   VIP           8490 non-null   object
 5   RoomService   8512 non-null   float64
 6   FoodCourt     8510 non-null   float64
 7   ShoppingMall  8485 non-null   float64
 8   Spa           8510 non-null   float64
 9   VRDeck        8505 non-null   float64
 10  Transported   8693 non-null   bool
dtypes: bool(1), float64(6), object(4)
memory usage: 687.8+ KB
```

This is further tailed by the inclusion of the describe method which basically places the columns side-by-side with all the required numerical data which provides a better understanding into the dataset.

	count	mean	std	min	25%	50%	75%	max
Age	8514.0	28.827930	14.489021	0.0	19.0	27.0	38.0	79.0
RoomService	8512.0	224.687617	666.717663	0.0	0.0	0.0	47.0	14327.0
FoodCourt	8510.0	458.077203	1611.489240	0.0	0.0	0.0	76.0	29813.0
ShoppingMall	8485.0	173.729169	604.696458	0.0	0.0	0.0	27.0	23492.0
Spa	8510.0	311.138778	1136.705535	0.0	0.0	0.0	59.0	22408.0
VRDeck	8505.0	304.854791	1145.717189	0.0	0.0	0.0	46.0	24133.0

This can also be depicted by using a correlation heatmap which would in addition to making the data more visually appealing, actually calculate the relationship between each column in the dataset.



Now it's the turn of one of the most undesirable elements in data analysis which is none other than ‘missing values’.Through the usage of pandas which possess one of the most highly sought after data analysis and manipulation libraries, the problem of missing values can be resolved much quicker as it would be highly awful to have them hanging without either fixing or eliminating them. After finding the amount of missing values per column, the total set missing values is divided into sets namely continuous and categorical where they are systematically dealt with through the usage of ‘mean’ which is used to fill the missing values with an aggregate or mean as implied and ‘value’ which is used to fill the missing values with a constant respectively.

Following this, the one-hot encoding method is used for one and only reason and that is to continuously feed labeled data to a variety of linear models, most notably the SVM which utilizes a standard kernel. The StandardScaler() method is one of the most coveted pre-processing methods that exist in the python language. Scaling is a very important step when it comes to tuning different algorithms with their datasets.The process of standardization usually occurs when the census is converted into 0s and 1s respectively.

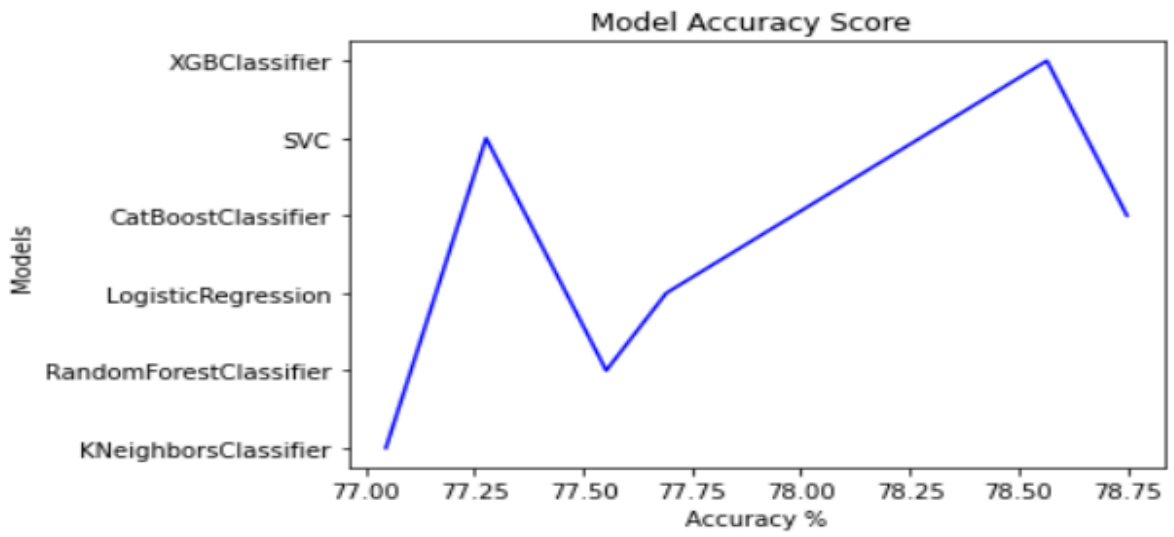
Legal, Ethical and Privacy Concerns

During the course of finding a solution to the problem statement none of the ethical boundaries have been breached as the entire dataset (train, test and sample submission) has been provided by kaggle which is the official organizer of the 'Spaceship Titanic' challenge.

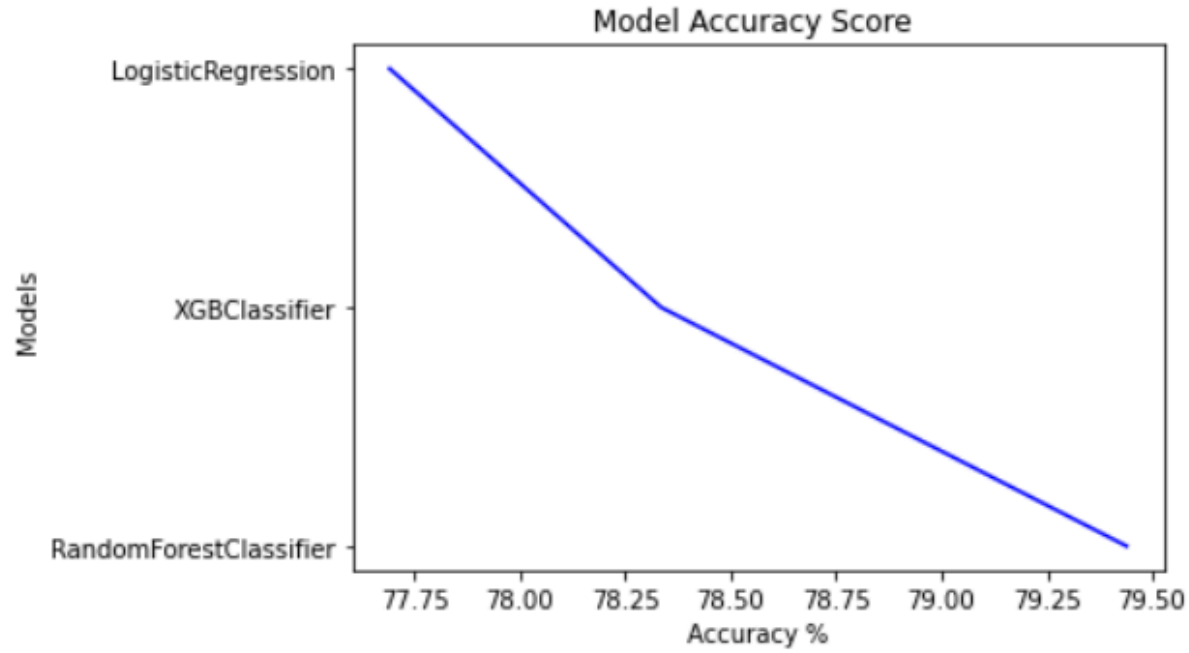
Critical Analysis

Classification is the right choice of algorithms in this case owing to the fact that the existence of labeled data helps in the prediction of the survival rate of the people at any given time. Labeled data according to the problem statement would infer the final two columns being PassengerId that refers to the identification number provided to the people prior to them boarding the ship and the target column Transported which would bear the prediction of either legitimacy or fallacy. Apart from this, a classification algorithm is used to determine fixed values such as True or False, Male or Female, whereas its counterpart regression is used to determine continuous values such as cost, profits, etc. This provides the justification for the aforementioned statement that classification is the right fit for this problem statement.

Coming to the choice of the algorithms used in the code, the majority of them are pitted against each other for the calculation of the highest accuracy without cross validation. Most notable ones included XGBoost which is based on gradient boosting which is a part of supervised learning algorithm that can predict a variable through the combination of a multitude of weaker models , Support Vector Machine is one of the many types of deep learning models which is capable of performing not only classification but also regression tasks, Catboost is algorithm almost similar to xgboost in its origins and finds its usage in a multitude of applications mostly search recommendation systems and autonomous cars, Logistic Regression has many similarities with linear regression with exception of manufacturing of predictions for categorical values, Random Forest Classifier is an ensemble learning algorithm that dabbles in both regression as well as classification, and the KNN Classifier is a non-parametric model as it lacks a real model and utilises the large amounts of training data as a model. After checking for accuracies, a model selection takes place and only the models which depict a high accuracy will be picked out from the lot and plotted on the graph as shown below.



As depicted in the above visualization, the XGBClassifier seems to be giving the best accuracy for the data that it has been trained on. Comparatively, the random forest classifier which usually has a reputation of providing high accuracy doesn't seem to be doing quite well. After this stage has been completed, cross validation is implemented with an ulterior motive in mind which is none other than trying to test the model's ability to predict by running it on new sets of data apart from trying to get rid of overfitting.

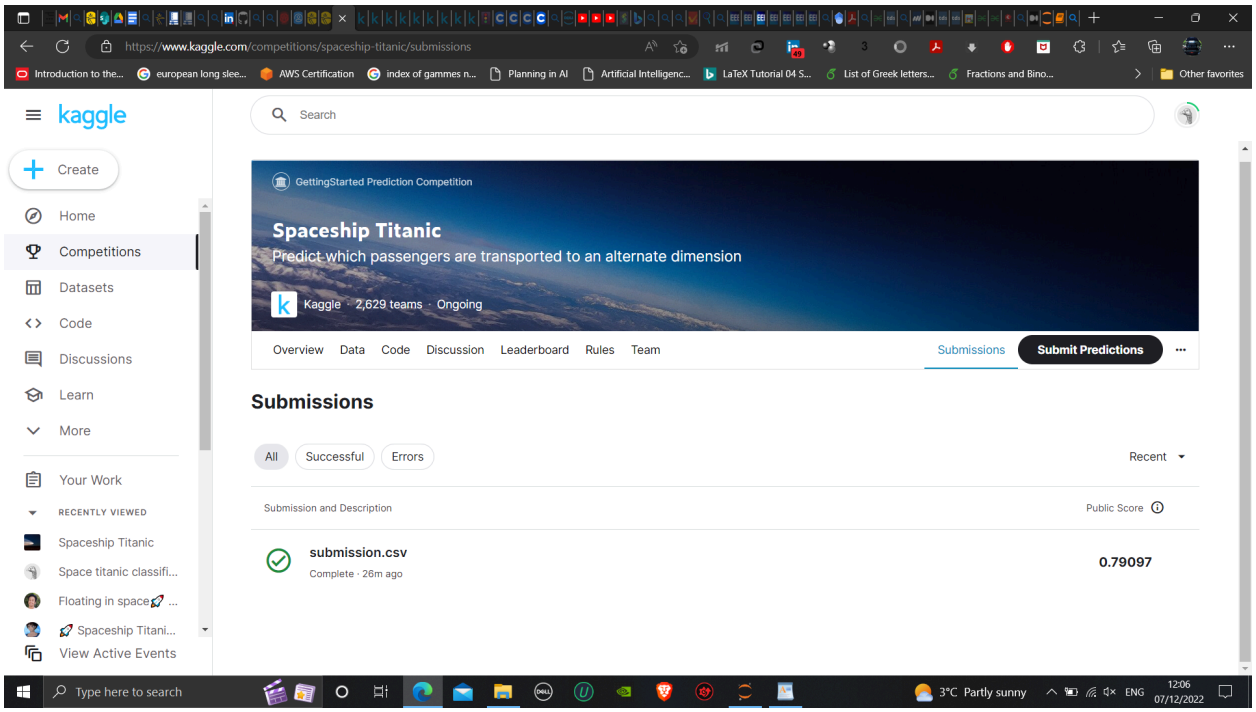


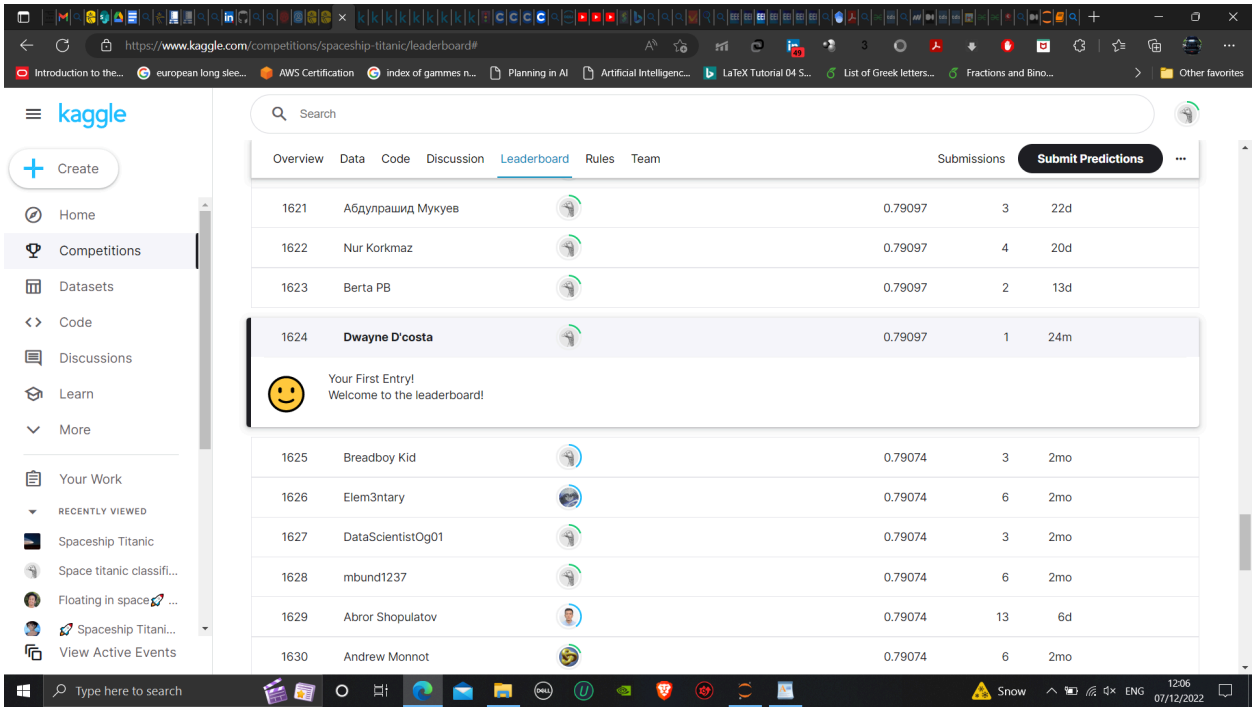
After cross validation has been implemented, it has been noticed that there is a

certain hike in the accuracies of the selected models with the random forest classifier gaining an edge over its counterparts as well as a hike in its own accuracy.

SID: 2050132

Kaggle Submission Proof





Result and Scope for Improvement

The final result that was obtained in the form of a ‘submission.csv’ file after the random forest classification model was chosen owing to its high accuracy from amongst its contenders. After uploading the final ‘submission.csv’ file to the kaggle challenge, I was awarded a full precision score of 0.79097.

The one way forward to improve the speed and accuracy of the models on a large scale would be through the inclusion of neural networks. Most common ones would include Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) which are extremely well versed in improving the performance of classification based prediction models.

References

1. Vinavatani, B., Panna, M.R., Singha, P.H. and Kathrine, G.J.W., 2022, May. AI for Detection of Missing Person. In 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 66-73). IEEE.
2. Shetty, J. and Pallavi, S., 2018, October. Predicting the Survival Rate of Titanic Disaster Using Machine Learning Approaches. In 2018 4th International Conference for Convergence in Technology (I2CT) (pp. 1-5). IEEE.
3. Ju, X., Yan, Z. and Wang, T., 2021, December. Overview of Optimization Algorithms for Large-scale Support Vector Machines. In 2021 International Conference on Data Mining Workshops (ICDMW) (pp. 909-916). IEEE.
4. Kumari, R. and Srivastava, S.K., 2017. Machine learning: A review on binary classification. International Journal of Computer Applications, 160(7).
5. Song, Q., Liu, X. and Yang, L., 2015, August. The random forest classifier applied in droplet fingerprint recognition. In 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) (pp. 722-726). IEEE.