**Project 2 : Classification**

**Titanic Dataset:**  <https://www.kaggle.com/competitions/titanic/data>

**Problem Statement:** Predication of the passenger survival in the titanic ship.

**Data understanding:**

|  |  |  |
| --- | --- | --- |
| Variable | Definition | Key |
| survival | Survival | 0 = No, 1 = Yes |
| pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Sex |  |
| Age | Age in years |  |
| sibsp | # of siblings / spouses aboard the Titanic |  |
| parch | # of parents / children aboard the Titanic |  |
| ticket | Ticket number |  |
| fare | Passenger fare |  |
| cabin | Cabin number |  |
| embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |

**Exploratory Data Analysis:**

1. Redundant variables: Passenger\_ID, Name, Ticket – As these columns have unique individual categorical feature. Going ahead and dropping it off.
2. Null value treatment:

* As the dataset, Cabin has more than 77% of null values. We can go ahead and drop the column.
* The column 'Age' can be imputed based on the median proportions based on gender sub class - Male and Female because of 20% data are null.
* Dropping off the rows from the columns 'Fare','Embarked' whose null values are less than 1%.

1. Categorical Encoding:

* From the dataset, as we have only 2 categorical variables – ‘Embarked’, ‘Gender’ and both are nominal dtype. Therefore, we can go ahead and one hot encoding.

1. Target value understanding:

From the dataset, we have 0 of 62.33% and 1 of 37.67% and proportions of two subclass are still good for the prediction and there no is no of using the smote technique.

**Feature engineering:**

1. **Correlation Matrix:**
2. We can see the variable like Embarked Q, EMbarked\_S are moderate co-relate each other.
3. The variables like Pclass and Fare are moderate co-relate each other.
4. The variable - 'Sex\_male' is highly co-relate with the target variable.
5. **Multicollinearity check:**

* From the analysis, we could observe that the variable 'Pclass' had the VIF value of 6.5 and after dropping it where all remaining variables have VIF threshold lesser than 5.

1. **Feature Engineering:**

From the forward and backward feature engineering techniques that features like

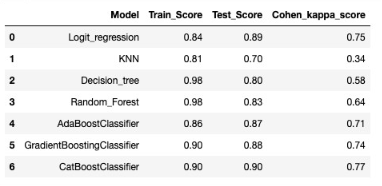
1. Sex\_male,
2. Embarked\_Q,
3. Embarked\_S,
4. Sibsp,
5. Parch are the key features having F1 score of 0.79

**Train test split:**

To do the base model, we need to train and test split. We have considering 70% of train data and 30% of test data with random state of 10.

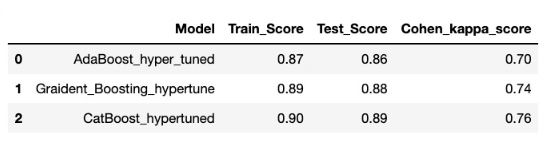
**Base Model:**

I’ve created individual nonlinear models and tested its training & test accuracies, the Cohen kappa score. Created report card with all of these scores against each individual models. Among them, the Cohen kappa score, accuracy scores of train and test are good for Adaboost, GraidentBoost and CatBoost. Therefore, going ahead and try hyper tuning for these models.



**Hyper Model Tuning:**

From the below three hypertuned models, we could see that CatBoost are outperforming in compare with other models. With the params of depth of 6, iterations of 500 and learning rate of 0.02. We could achieve the best Cohen kappa score of 0.76.

****

**Business Interpretation:**

According to this model, the class 0 is very sensitive in compare to class – 1. From the classification report, the derived F1score of class - 0 is 0.92 is higher than class- 1 is 0.85. We could strongly say that prediction of class- 0 (i.e., Survived “NO”) is 0.92% and only 0.08% of wrong prediction and whereas for class-1 (i.e., survived “Yes”) is 0.84% and only 0.16 of wrong prediction. Therefore, the model is recommended for the deployment.

**Limitations:**

1. Need more records to carry out for better results.
2. Features are very limited.