1&1 Data Science Challenge- Titanic

In this repository, you will find the road map that I followed to build a machine learning model for 1&1 Data Science challenge using the famous titanic dataset from kaggle. The model's predictions have 76.5% accuracy score.

In the building of the solution, I did the following steps:

A. Introduction (importing necessary libraries)
B. Loading datasets
C. Exploratory data analysis
    1. Understanding data
    2. Target variable analysis (what values it has, if there are missing values)
    3. Features analysis - because there are relatively few features, I chose to explore each one, finding interesting information about them.
        a. I found out that passenger in first class have higher chance of survival.
        b. At the first glance, the 'Name' feature seemed to not help us, because it have a different value for every passenger, and therefore, the predictive model might not differentiate between passenger that survived or not based on their name, but, I saw that passengers have also title in their name (Mr., Mrs., Miss. etc.). By transforming this feature, we can have a better model.
        c. Females have more chances to survive.
        d. Because the 'Ticket' feature has both numerical and categorical values and does not follow a visible pattern, I decided to remove this feature.
        e. Because the ‘Cabin’ feature has a lot of missing values, I decide to drop this feature.
        f. Passengers embarked in C (Cherbourg) have higher chances of survival.
        g. When looking at numerical features, although the correlation coefficients are not high, subpopulations in these features can be correlated with the survival.
        h. ‘Age’ feature has missing values. Also, the distribution look normal and it seems that younger passengers have survived less.
        i. The number of passengers that were alone on the ship is higher that the number of passengers that went with a member of family. It also looking like there is a relationship between family and chances of survival. Because these two features are related to family, they can be combined.
        j. The distribution of ‘Fare’ feature is skewed, with the majority of values being low.
D. Feature engineering
    1. Missing values imputation (imputing missing values in ‘Age’ feature with median value, imputing missing values in ‘Embarked’ feature with most common value, and dropping ‘Cabin’ feature because of high number of missing values).
    2. Feature transformation
        a. Extracting ‘Title’ from ‘Name’ feature.

b. Transform values of ‘Sex’ feature in numerical values.
c. Combine ‘SibSp’ and ‘Parch’ features into a single ‘Family size’ feature.
d. Drop unnecessary features ('PassengerId', 'Ticket', 'SibSp', 'Parch', ‘Name’).

3. Categorical encoding – encoding categorical features using OneHotEncoder.

E. Model development
1. Select independent variables and target separately.
2. Split train set in X_train, X_test, y_train, y_test, with 10% of train data for evaluation.
3. Trying different algorithms (Random Forest, LightGBM, XGBoost). For each algorithm, metrics like accuracy_score, precission, recall, confusion_matrix and auc_score were calculated. I found out that LightGBM have the best results (86.6% accuracy score for a baseline model).
4. Hyperparameter tuning for LightGBM to find the best combination using GridSearchCV. Best model have the following hyperparameters: n_estimators = 100, max_depth=4, max_leaf_nodes=2, min_samples_leaf=3, min_samples_split=3. Accuracy score = 87.7%.

F. Model explainability
1. Feature importance
a. We can see that the most important features are: Title_Mr, Fare, Age, Pclass_3, FamilySize.
b. Some of the features created in the Feature Transformation part are found in top (Title_Mr, FamilySize, Pclass_3).
c. We can see that these features have a higher impact on survival.
2. Shap
a. By looking at the shap plot, we can see how each feature contribute to the target.
b. Blue color shows low values of feature, while red color shows high values of features. Also, points found in the left of vertical line at 0 shows a lower probability for that observations to survive, while points found at right shows a higher probability to survive.
c. Therefore, passenger that have not ‘Mr.’ title (Title_Mr = 0) are more likely to survive.
d. Passenger that are not found in ticket class 3 (Pclass_3 = 0) are more likely to survive.
e. Passengers with a high number of family members have low chances to survive.

G. Submission – preparing csv file to submit.