PREDICTING CANCELLED TENDERS

NAJWA ALMULHEM

ABSTRACT

The goal of this project is to build a binary classification model for predicting cancelled tenders using tenders' numerical and categorical features. The data was provided from The Electronic Government Procurement System in Saudi Arabia .Data was masked for privacy protection. The features were reengineered and scaled to be used by different algorithms like decision tree, logistic regression xgboot & k neighbors. Then choosing the decision tree model as a binary classifier and fine tune it to achieve the best results. This will also create a better understanding of the tenders that were cancelled, and which features are important for cancelled tenders.

DESIGN

This project is part of Tuwaiq Four Week Bootcamp. The data was provided from The Electronic Government Procurement System in Saudi Arabia and has binary class for status "Completed" or "Cancelled". Predicting cancelled tenders might save time & money of both public tender workers as well as suppliers & contractors.

DATA

This data has 160732 rows with 5 categorical features like city invitation type, tender type and activity name. 3 date features for creation, submission and last offer presentation .In addition to 4 numerical features like conditions booklet price ,estimated value , number of regions & number of categories

ALGORITHMS:

FEATURE ENGENNERING:

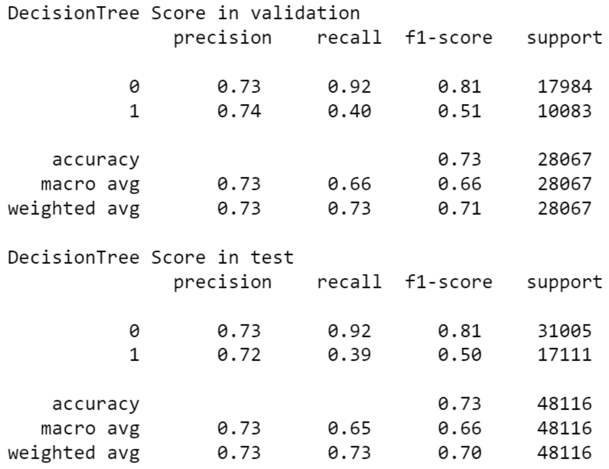
- Converting regions fields into columns since the region filed can contain multiple regions or cities, we will cover those values to columns to capture it as a feature One hot encoding for categorical fields
- Removing dates as it does not affect cancellation.
- Convert Invitation count, estimated value & condition booklet price into buckets or bins to increase accuracy.
- Do one-hot encoding for categorical variables.

Models:

A model of the mean was set as a base model. Logistic regression ,decision trees, xgboost & KNeighbors were built and using 5 folds cross validation against train dataset to evaluate the models based on recall since our focus is predicting cancelled tenders. Decision trees has the highest mean score for recall since we care about predicting cancelled tenders.

Model	Accuracy	Recall	F1
Logestic Regression	0.646	0.626	0.622
Decision Tree	0.733	0.657	0.664
KNeighbors	0.669	0.643	0.648
XGBClassifier	0.768	0.646	0.649

Then Using RandomizedSearchCV to fine tune the Decision tree model and the recall score was 0.66 on test data and validation data.



```
DecisionTree Score in validation
              precision    recall  f1-score   support

           0       0.73      0.92      0.81     17984
           1       0.74      0.40      0.51     10083

    accuracy                           0.73     28067
   macro avg       0.73      0.66      0.66     28067
weighted avg       0.73      0.73      0.71     28067

DecisionTree Score in test
              precision    recall  f1-score   support

           0       0.73      0.92      0.81     31005
           1       0.72      0.39      0.50     17111

    accuracy                           0.73     48116
   macro avg       0.73      0.65      0.66     48116
weighted avg       0.73      0.73      0.70     48116
```

TOOLS

- *Sklearn*
- seaborn
- PIL
- wordcloud
- matplotlib

COMMUINCATION

A Presentation was submitted with this report.