HaikuJAM Data Science Case study Report

Submitted by

Manoj Kumar Ainala

HaikuJAM Data Science Case study Report

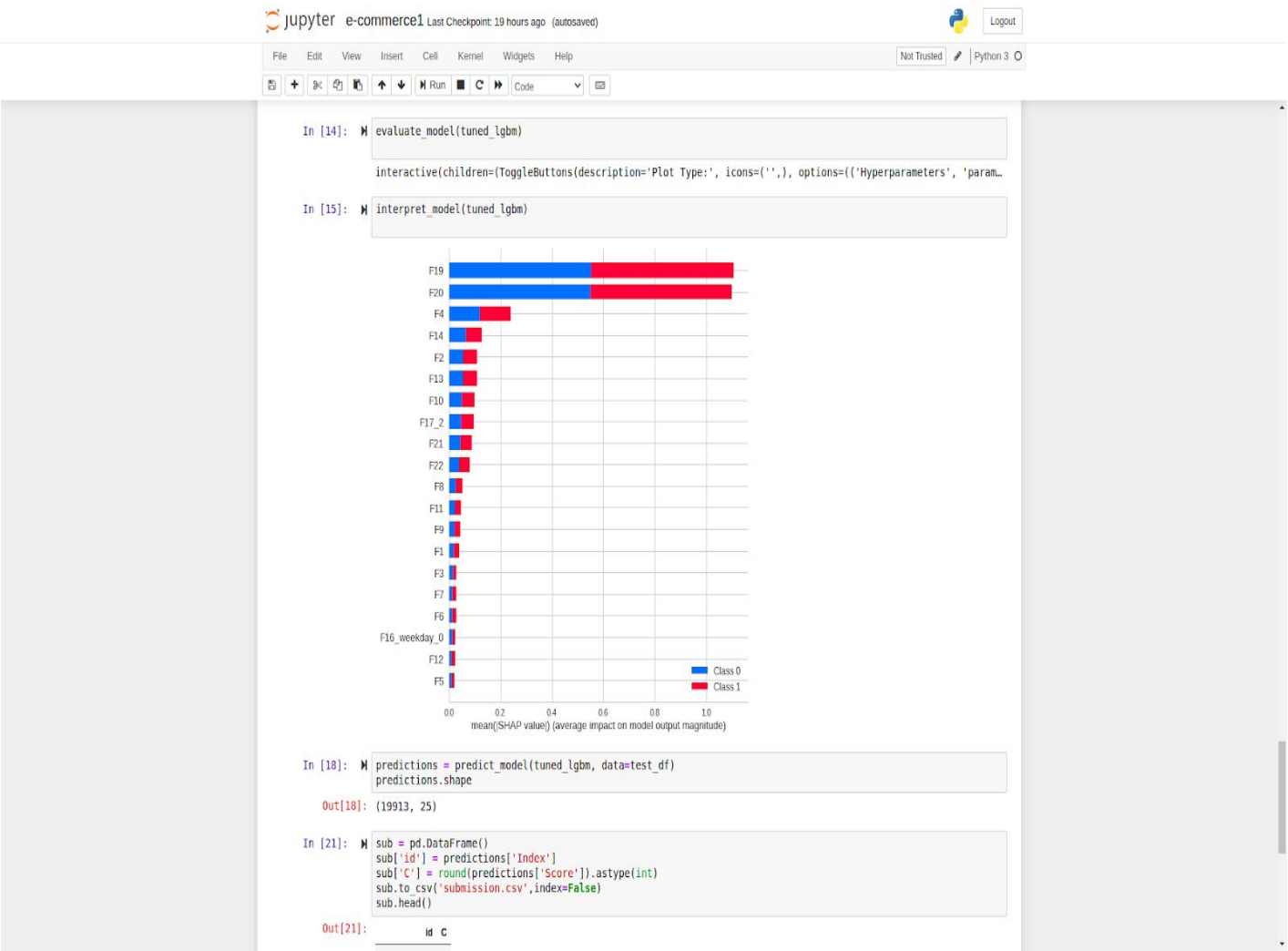
OverView: The aim of this case study is to predict whether the user buys the item or not. By Observing the Dataset we can say this is a classification problem and that too its a binary classification. By using the Machine Learning Models Like SVM, Random Forest, Decision Tree, XGboost,LightGBM, etc., I have solved this problem.

Data Description: The Data set contains two CSV files i.e., test data and train data. There are 23 attributes in the test data and 24 attributes in the train data. The final 24th attribute in the train data is the required solution for the case study.
While training the model I had changed the F15, F16 columns, Because those two attributes are defining date and are in the form of mm/dd/yyyy. For this I had used label encoding where these formats changed into numerical values. As the machine understands only the numerical values then it's better to change all the columns into integers and some of the data is negative, so I have changed that to zeros. As, the data is not big, So, I haven't done any feature engineering or feature selection.

Models: Firstly I used traditional Machine Learning models vizly., SVM, Random Forest, XGBoost, LightGBM etc., but it took a lot of time to compile and return the results. Later I used PyCaret to get the results. Pycrate is an open source, low-code machine learning library in Python which gives the results within seconds and all the traditional models are pre defined in Pycaret. So it is very easy to get to results and as well as this gives the comparison of different models based on the evaluation metrics like Accuracy, AUC, F1 score etc., .

Results: The Highest Accuracy was 76.19% using LGBM technique. By plotting the graph the most important features were F19, F20.

Important Features:



Comparison of Models:

35	Polynomial Threshold	None
36	Group Features	False
37	Feature Selection	False
38	Features Selection Threshold	None
39	Feature Interaction	False
40	Feature Ratio	False
41	Interaction Threshold	None
42	Fix Imbalance	False
43	Fix Imbalance Method	SMOTE

Type *Markdown* and LaTeX: α^2

In [4]:
```
compare_models()
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
0	Light Gradient Boosting Machine	0.7619	0.7273	0.0520	0.7006	0.0967	0.0650	0.1438	0.8200
1	CatBoost Classifier	0.7595	0.7216	0.0719	0.5790	0.1279	0.0780	0.1374	17.2703
2	Gradient Boosting Classifier	0.7559	0.7182	0.0104	0.6601	0.0205	0.0130	0.0597	29.2452
3	Naive Bayes	0.7546	0.5069	0.0000	0.0000	0.0000	0.0000	0.0000	0.0686
4	Ridge Classifier	0.7546	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1364
5	Linear Discriminant Analysis	0.7546	0.6864	0.0000	0.0000	0.0000	0.0000	0.0000	0.4240
6	Ada Boost Classifier	0.7544	0.6913	0.0050	0.4646	0.0100	0.0046	0.0257	5.7161
7	Quadratic Discriminant Analysis	0.7535	0.5001	0.0023	0.2426	0.0044	0.0001	0.0021	0.2199
8	Extra Trees Classifier	0.7534	0.6871	0.0094	0.3939	0.0183	0.0070	0.0265	3.5004
9	Extreme Gradient Boosting	0.7534	0.7183	0.1136	0.4897	0.1844	0.1013	0.1394	16.7909
10	Random Forest Classifier	0.7356	0.6699	0.1088	0.3689	0.1680	0.0633	0.0801	0.5221
11	K Neighbors Classifier	0.7061	0.5121	0.1091	0.2620	0.1540	0.0115	0.0131	0.4476
12	Decision Tree Classifier	0.6676	0.5588	0.3450	0.3302	0.3374	0.1157	0.1158	1.9269
13	Logistic Regression	0.5049	0.5061	0.5078	0.2497	0.3348	0.0087	0.0101	0.0875
14	SVM - Linear Kernel	0.5001	0.0000	0.4919	0.2434	0.3256	-0.0039	-0.0045	6.3105

Out[4]:
```
LGBMClassifier(boosting_type='gbdt', class_weight=None, colsample_bytree=1.0,
               importance_type='split', learning_rate=0.1, max_depth=-1,
               min_child_samples=20, min_child_weight=0.001, min_split_gain=0.0,
               n_estimators=100, n_jobs=-1, num_leaves=31, objective=None,
               random_state=123, reg_alpha=0.0, reg_lambda=0.0, silent=True,
               subsample=1.0, subsample_for_bin=200000, subsample_freq=0)
```