Solutions to the questions mentioned in Challenge pdf

1)The duration attribute played a major role in determining the deposit for a given customer.I made the use of LOFO algorithm to predict this attribute,The reasons are as follows:-

LOFO (Leave One Feature Out) Importance calculates the importances of a set of features based on a metric of choice, for a model of choice, by iteratively removing each feature from the set, and evaluating the performance of the model, with a validation scheme of choice, based on the chosen metric.

LOFO first evaluates the performance of the model with all the input features included, then iteratively removes one feature at a time, retrains the model, and evaluates its performance on a validation set. The mean and standard deviation (across the folds) of the importance of each feature is then reported.

LOFO has several advantages compared to other importance types:

It does not favor granular features
It generalises well to unseen test sets
It is model agnostic
It gives negative importance to features that hurt performance upon inclusion
It can group the features. Especially useful for high dimensional features like TFIDF or OHE features.
It can automatically group highly correlated features to avoid underestimating their importance.

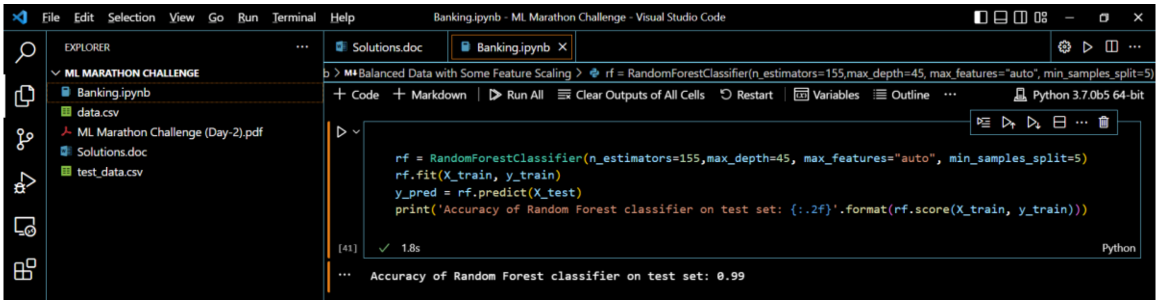
apart from duration,pdays,month and poutcome are some other features which had a great impact.

2)Using Logistic Regression is the best bet here.As Logistic Regression is the fastest and versatile algorithm which can used to solve such classification problems in supervised learning.

3)Now a low variance and low bias is considered as an idiol dataset for a ML engineering,however it is impossible to have such dataset ie if we try to decrease bias the variance increasing.But in case this happens it will mean that the model has data very closely related or it is imbalanced dataset with some values missing.

4)On of the ways of missing values is to drop the tuples if the dataset is very large.Otherwise if the data is numeric we can replace the missing values with mean, mode or median.If even after manipulating the dataset the classification algorithm is not working we can make use of KNN(Nearest Neighbors Imputations) Missing values are imputed using the k-Nearest Neighbors approach where a Euclidean distance is used to find the nearest neighbors.

5)If we dont split the dataset and use the same dataset than it will lead to developing a model with better accurary which may or may not make good prediction as it will give false impressions.



File
Edit
Selection
View
Go
Run
Terminal
Help
Banking.ipynb - ML Marathon Challenge - Visual Studio Code
EXPLORER
ML MARATHON CHALLENGE
Banking.ipynb
data.csv
ML Marathon Challenge (Day-2).pdf
Solutions.doc
test_data.csv
Solutions.doc
Banking.ipynb
b > M↓Balanced Data with Some Feature Scaling > rf = RandomForestClassifier(n_estimators=155,max_depth=45, max_features="auto", min_samples_split=5)
+ Code
+ Markdown
Run All
Clear Outputs of All Cells
Restart
Variables
Outline
Python 3.7.0b5 64-bit
[41]
1.8s
Python

```
rf = RandomForestClassifier(n_estimators=155,max_depth=45, max_features="auto", min_samples_split=5)
rf.fit(X_train, y_train)
y_pred = rf.predict(X_test)
print('Accuracy of Random Forest classifier on test set: {:.2f}'.format(rf.score(X_train, y_train)))
```

```
Accuracy of Random Forest classifier on test set: 0.99
```