Assignment:

* I have done some preprocessing on the dataset like removing rows having Null values for continuous features. There were around 1000 rows, which could be imputed but in interest of time it was less than 1% of data, so I removed them.
* I created one hot vectors for categorical variables.
* I have used SelectKBest approach to shortlist good variables using validation set.
* I have also tried to see optimal threshold using ROC Curve.
* I used Random Forest Classifier model for the whole exercise. In interest of time, I didn’t try many algorithms.
* I also removed highly correlated features.

If there was more time,

* I would like to include persisting and loading methods for my Estimators as that would not require synchronous training and predicting instance.
* I would also like to try more algorithms and do hyper parameter tuning.

I tried to keep engineering infrastructure practicality in mind while writing the code and used docker files, Scikit-learn Estimator style methods for easy persisting and loading of estimator methods. But further, I would have to develop unit test cases for each estimator. Also, develop integration tests for feature testing. I would need to build my own docker file with fixed package versions. Also create docker-compose service to host this service and create endpoints for the UI to integrate with my model.