

Micro-Credit Defaulter Model

Submitted by:

Parija Deshmukh

**INTRODUCTION**

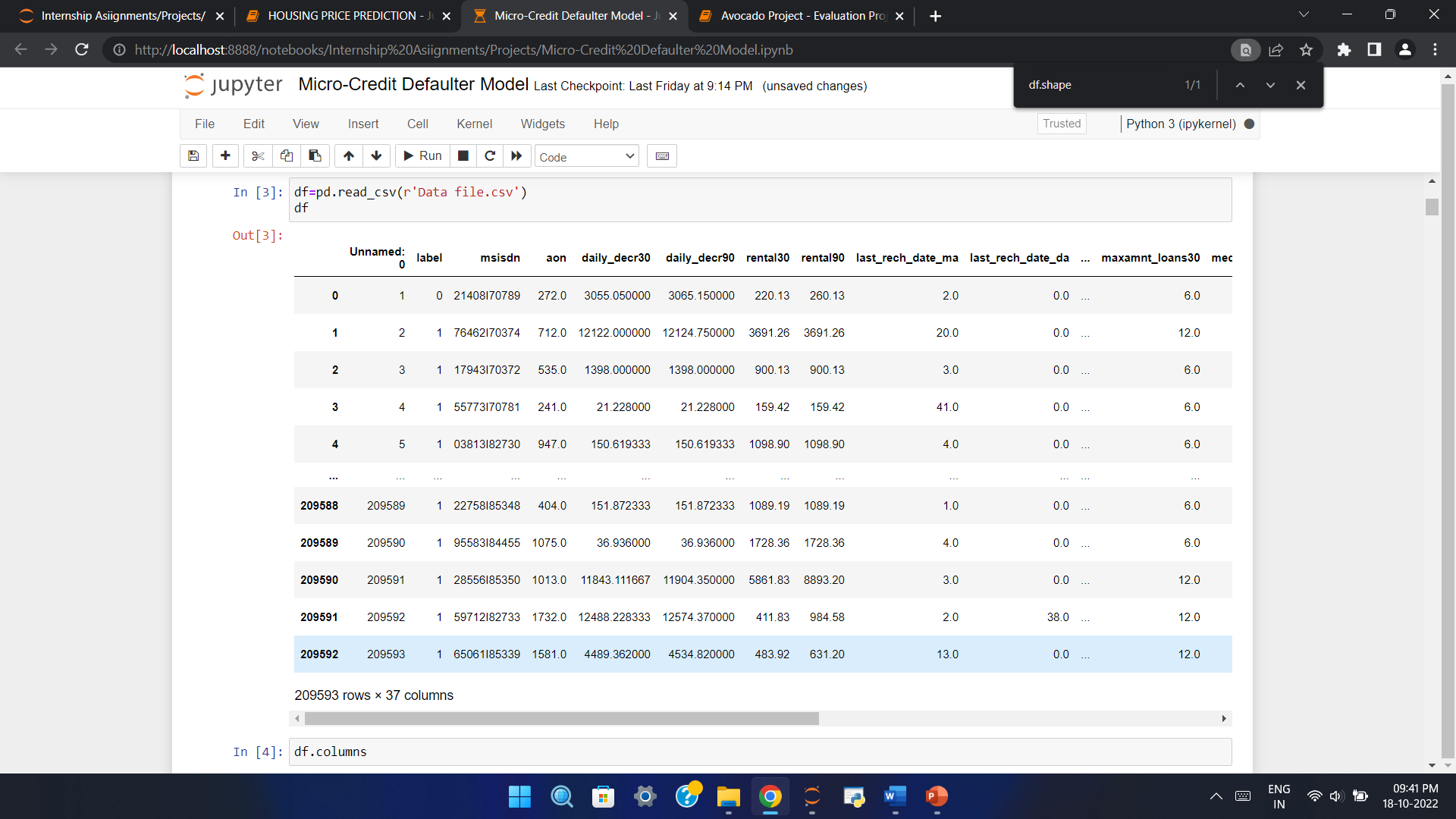
A Microfinance Institution (MFI) is an organization that offers financial services to low-income populations. Microfinance services (MFS) becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The MFS provided by MFI are Group Loans, Agricultural Loans, and Individual Business Loans and so on.

Now a days telecom industries launched various products and have developed its business and organization based on the budget operator model, offering better products at Lower Prices to all value conscious customers through a strategy of disruptive innovation that focuses on the subscriber.

They understand the importance of communication and how it affects a person’s life, thus focusing on providing their services and products to low-income families and poor customers that can help them in the need of hour. They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in some days e.g., in 5/10/15 days.

**Analytical Problem Framing**

* Data Sources and their formats



Dataset is huge with over 2 lakhs of rows and 37 columns. Data set has different features like age, mobile number, average main account balance for 30 and 90 days, data related to account recharge, number of loans taken and other few data related to loan. There is target columns as Label having categorical values “0” for loan not paid and “1” for loan paid.

* Data Pre-processing Done

1. Few irrelevant data for column maxamnt\_loans30 has been dropped.
2. Checked null values and no null values found.
3. Deleted Unnamed: 0, pcircle columns as it wasn’t significantly contributing in data analysis.
4. Converted object type columns into numeric columns using LabelEncoder.
5. Target column- Label was imbalanced hence balanced the column with SMOTE technique.

* Data Inputs- Logic- Output Relationships

1. Frequency of data account recharged in last 90 days is highly negatively correlated with target column – Label with correlation of -0.005501 value.
2. cnt\_ma\_rech30 and cnt\_ma\_rech90(Number of times main account got recharged in last 30/90 days) is highly positively correlated to Label with value: 0.237201 and 0.236295
3. fr\_da\_rech30 is least correlated with Label means it doesn’t impact much on output.

* Hardware and Software Requirements and Tools Used

The preliminary step involved in devising a model is loading the required libraries. In this case, we mainly load four libraries namely pandas, numpy, scipy, sklearn and seaborn.

1. Pandas: Pandas is a python package which is quite quick, easy to use and structured in nature. Pandas data-frame is mainly used for data analysis purposes. Pandas also helps us to handle missing data, data to be reshaped and data transformation methodologies.
2. Numpy: Numpy is essentially used for creating very powerful and intuitive n-dimensional arrays. It offers various mathematical functions and also supports various kinds of computing hardware and software requirements. It is an open-source project and also contains various array-objects which are generally quick for usage.
3. Sci-kit: Sci-Kit is one of the most efficient and useful machine learning libraries in python. It provides various statistical and mathematical tools. It also has various techniques like regression, clustering etc. in it.
4. Seaborn: Seaborn helps us to create plots and live-interactive statistical plots and images. It also contains matplotlib which helps us to view our data in a much more intuitive manner.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

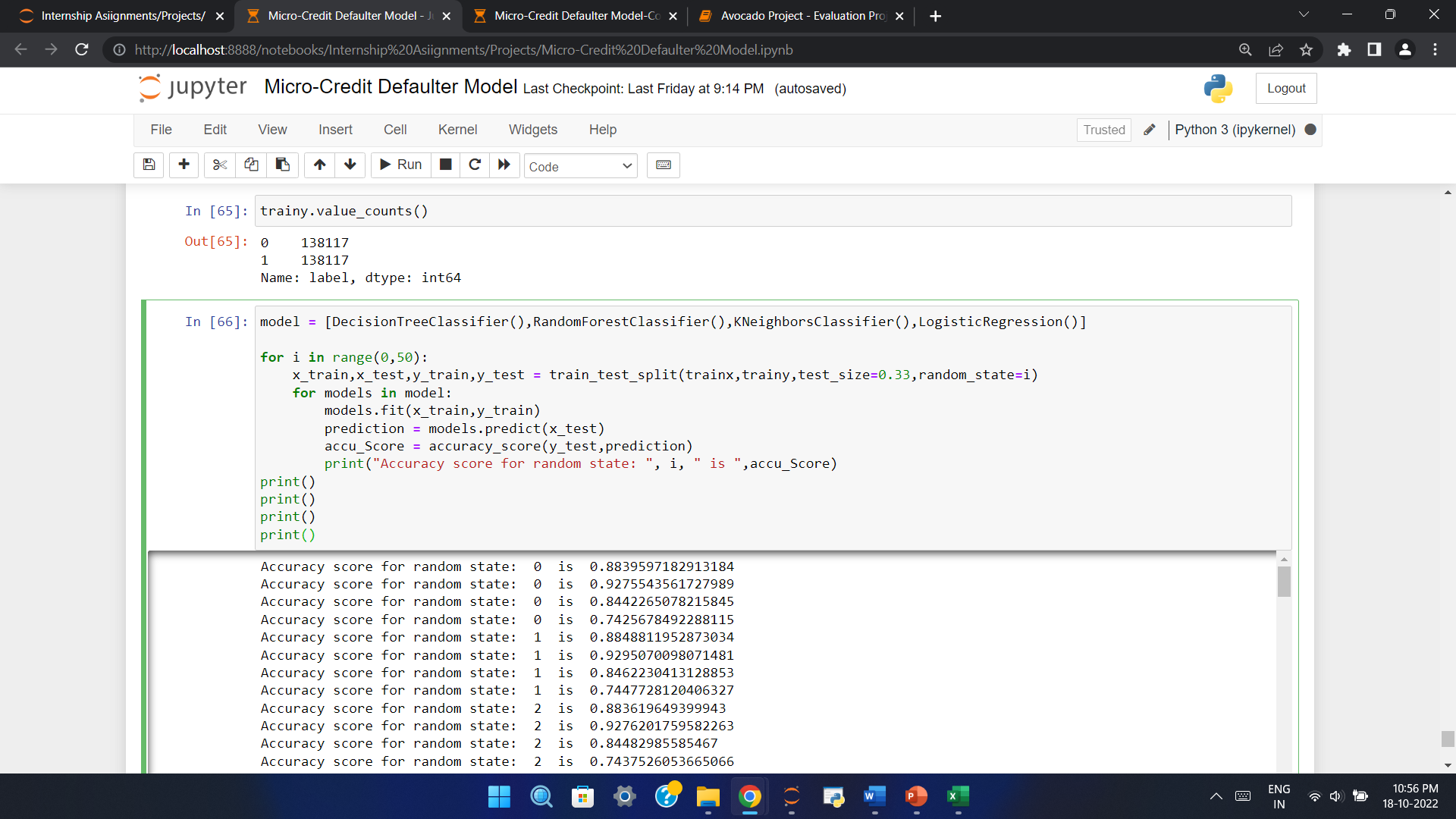
As target column is categorical with two values (0 and 1) we can select Logistic regression or classifier models.

In this case we have used Decision Tree Classifier, Random Forest Classifier, KNeighbours Classifier and Logistic Regression.

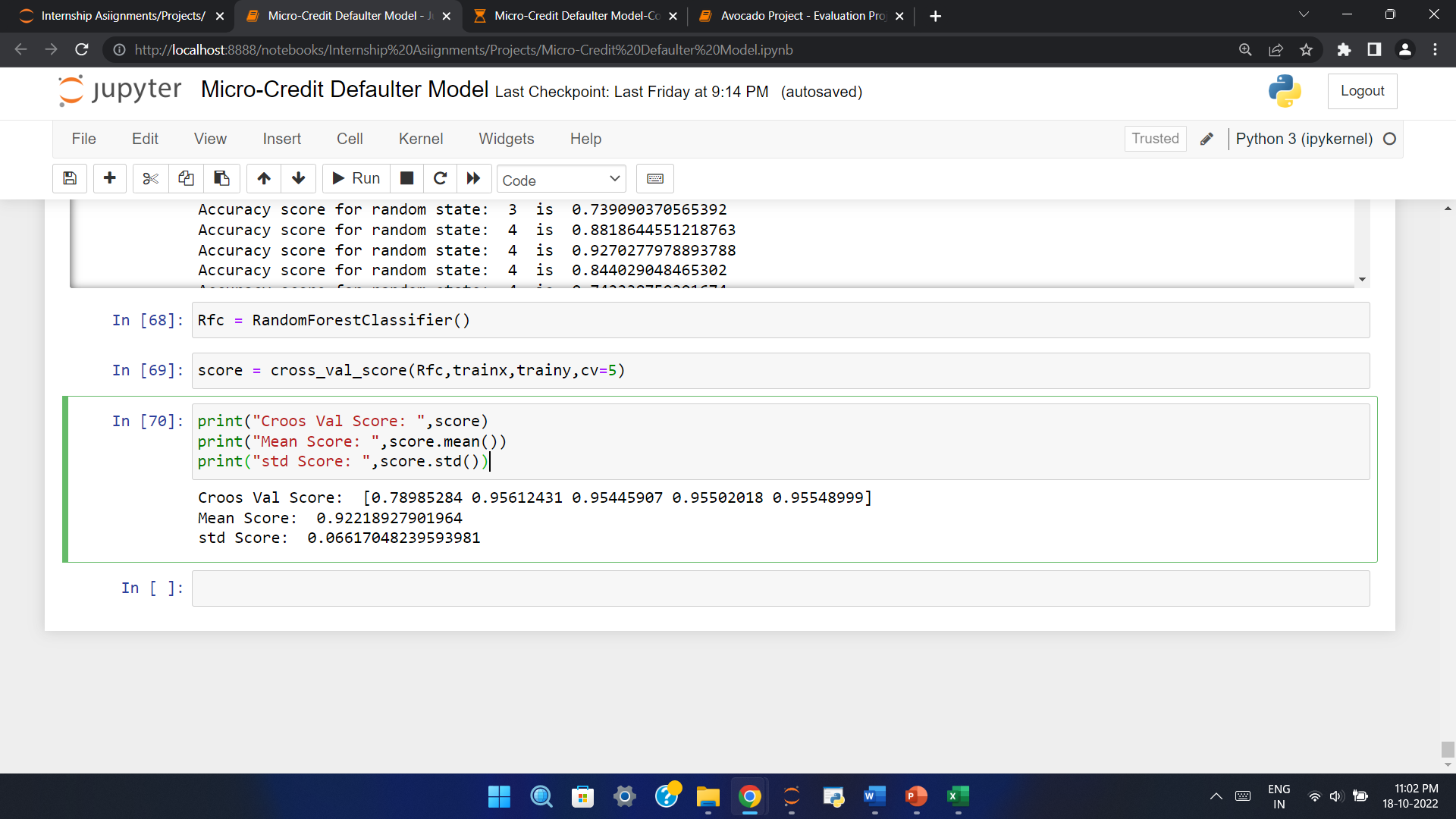
* Testing of Identified Approaches (Algorithms)

1. Decision Tree Classifier
2. Random Forest Classifier
3. KNeighbours Classifier
4. Logistic Regression

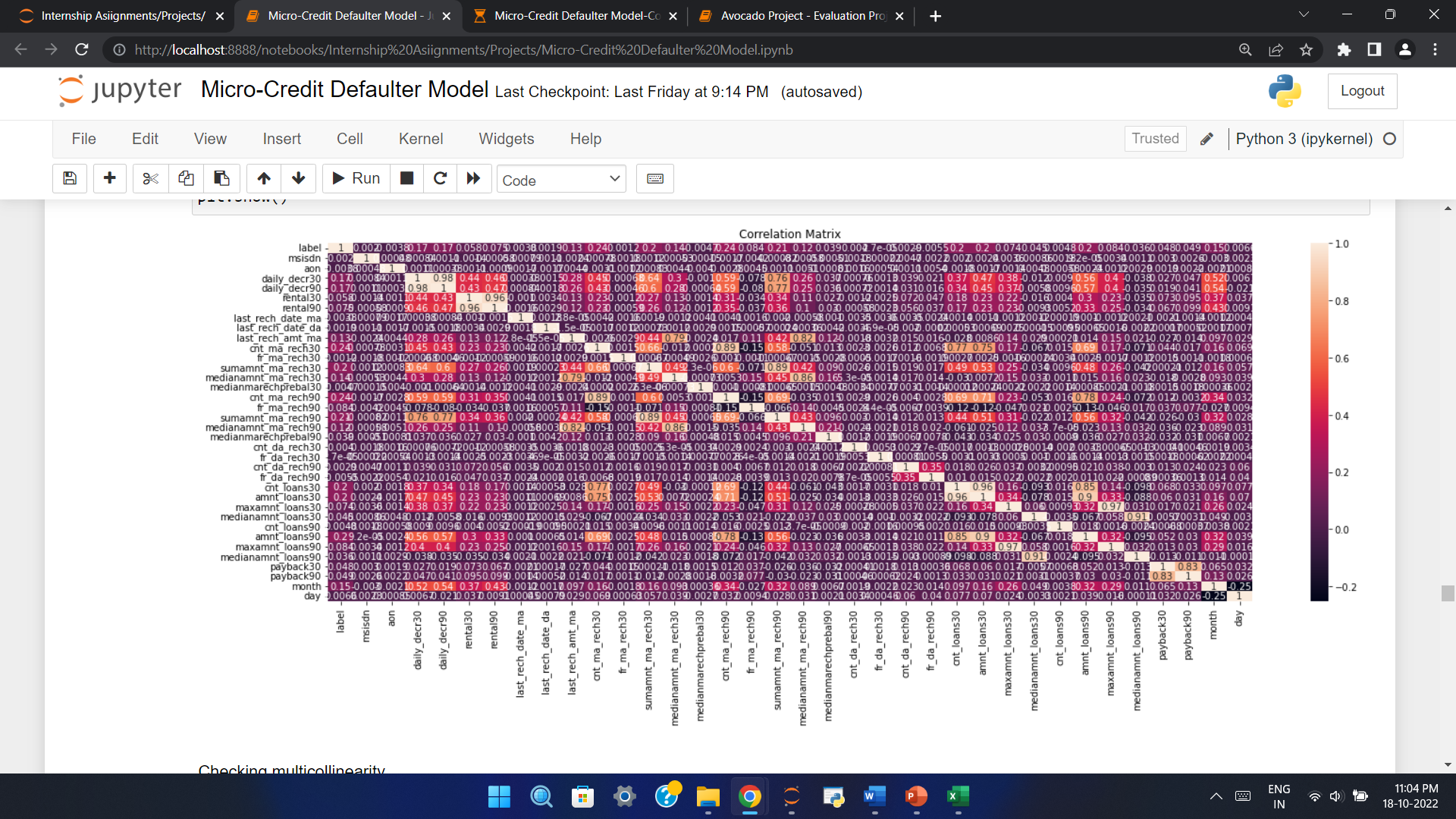
* Run and evaluate selected models



* Key Metrics for success in solving problem under consideration



* Visualizations



* Interpretation of the Results

1. Random Forest model is giving best accuracy score : 92%
2. Std is 0.06 and mean accuracy is 93%.

**CONCLUSION**

Random Forest Model is the best model out of all model tested above and by looking this we can conclude that our model is predicting around 93% of correct results for Label.