**Title - Banking domain Credit Card department’s data prediction**

**Scope: Risk analysis**

Types of data - Structured, Semi-strutted data

**Machine Learning Algorithms:**

* **Target column with Distributional values**
* **Target column with binary values**

**Technology stack –**

* Python
* ML algorithms
* SQLite (Structured source)
* CSV, TSV, Multi-delimited values, JSON ( Semi-structured sources)

Project Duration:

|  |  |
| --- | --- |
| Efforts in Weeks | Tasks |
| 1 | Requirement analysis ,generate sample data ,  desired output - eligible or not |
| 2 | - python integration with source systems (SQLite, CSV, TSV, Multi-delimited, JSON) and data cleaning |
| 1 | Algorithm implementation and model training |
| 1 | Test algorithm results and check the error |
| N/A | Git integration |
| 1 | Performance tuning |
| 1 | Documentation |

**CSV dataset description:**

Dataset contains 26052 records with 7 columns (fields)

The description of columns (fields) :

* City column has 986 distinct values
* Date column has 20 months data
* Card Type column contains 4 types of categories
* Exp Type (expenditure type) column contains 6 types of values: Bills, Food, Entertainment, Grocery, Fuel, and Travel
* Gender column
* Amount column: contains transaction amount
* EMI Paid: This column contains binary value about whether user paid EMI by due date or not

Note: There are many other columns like transcation\_id, card\_number, etc. which are already filtered by the data source team.

Tasks:

1. Clean dataset (if applicable)
2. You have to visualize data , based on gender and their transaction count
3. You have to visualize data , based on gender and sum of their transactions amount
4. Find out the top 5 cities based on defaulters count (who didn’t pay EMI by due date) display result in descending order based on defaulters count.
5. Find out the top 5 cities based on count of “EMI paid” (who paid EMI by due date) in descending order based on count.
6. Split dataset in test and train sets.
7. Use EMI\_Paid column as target column, apply KNN and generate confusion matrix
8. Use EMI\_Paid column as target column, apply Naive bayes and generate confusion matrix