**Integrated Capstone Project**

**This Case Study has three checkpoints defined in it.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Check Point Topics** | **Remarks** | **Max Marks** | **Faculty** |
| 1.1 SQL (30 marks)  1.2Informatica (ETL)  (40 Marks)  1.3 Power-BI (30 Marks) | **Business Reporting** | **100** | **Hemant & Arif** |
| 2.1 Exploratory Data Analysis (50 marks)  2.2 Statistical Analysis  (50 marks) | **Business Analysis** | **100** | **Manish** |
| 3. Data Engineering (Big Data) | **Data Engineering** | **100** | **Praveen** |
| 4 Model Building using ML algorithms | **Data Science** | **100** | **Manish** |
| **TOTAL =** | | **400** |  |

**Domain: BFSI**

**About:**

The Data is related to credit product details of masked customers. The data is extracted from various sources, which is not in proper format. Total 6 files consist of demographic of Dev and oot

**What is Expected?**

The motive here is to build a credit risk scorecard for affordable home loan applicants and help in approving / declining customer applications based on the same.

Hence the model should provide a probability score that gives the likelihood that applicant will default (will not pay his loan) if approved.

Affordable home loan scheme in Axis Bank is called HL ASHA , where sanction limit is upto 35 lakhs.

**Data Dictionary:**

**Bureau Variables**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| AL\_open\_TL | Auto Loans Active Tradeline |
| AL\_No\_of\_TL | Auto Loans No. of tradeline |
| All\_open\_TL | All Active Tradeline |
| All\_Delq\_90\_plus\_TL\_24\_M | All Delinquency( 30+ DPD) Trade Lines in last 24m |
| All\_No\_of\_TL | All No. of tradeline |
| BL\_open\_TL | Business Loans Active Tradeline |
| BL\_No\_of\_TL | Business Loans No. of tradeline |
| CC\_open\_TL | Credit Card Active Tradeline |
| CC\_latest\_uti | Credit Card Latest utilisation |
| CC\_No\_of\_TL | Credit Card No. of tradeline |
| CL\_open\_TL | Consumer Loans Active Tradeline |
| GL\_open\_TL | Gold Loans Active Tradeline |
| GL\_No\_of\_TL | Gold Loans No. of tradeline |
| Latest\_Bureau\_prd | Latest Bureau Product |
| ALL\_Alg\_max\_delq\_24m | Max Delinquency in all trades in last 24m |
| CL\_Alg\_max\_delq\_24m | Max Delinquency in Consumer Loans trades in last 24m |
| HL\_Alg\_max\_delq\_12m | Max Delinquency in Home Loan trades in last 12m |
| HL\_Alg\_max\_delq\_24m | Max Delinquency in Home Loan trades in last 24m |
| PL\_Alg\_max\_delq\_24m | Max Delinquency in Personal Loan trades in last 24m |
| Sec\_C\_Alg\_max\_delq\_24m | Max Delinquency in Auto Loans and Home Loans trades in last 24m |
| Sec\_O\_Alg\_max\_delq\_24m | Max Delinquency in Products other than Auto Loans anf Home Loans trades in last 24m |
| USEC\_C\_Alg\_max\_delq\_24m | Max Delinquency in Credit card and Personal Loans trades in last 24m |
| USEC\_O\_Alg\_max\_delq\_24m | Max Delinquency in unsecured products other than Credit Card and Personal Loans trades in last 24m |
| All\_min\_Vin\_bureau | Months Since last opened last tradeline |
| No\_of\_enq\_last\_9m\_All | No. of All Enq in last 9m |
| No\_of\_enq\_last\_12m\_CL | No. of Consumer Loans Enq in last 12m |
| No\_of\_enq\_last\_24m\_PL | No. of Personal Loan Enq in last 24m |
| PL\_open\_TL | Personal Loan Active Tradeline |
| PL\_No\_of\_TL | Personal Loan No. of tradeline |
| SEC\_C\_Delq\_TL\_24\_M | Auto Loans and Home Loans Delinquency( 0+ DPD) Trade Lines in last 24m |
| TWL\_open\_TL | Two Wheeler Loans Active Tradeline |
| TWL\_No\_of\_TL | Two Wheeler Loans No. of tradeline |
| USEC\_C\_TL\_ENQ\_RATIO | Unsecured Core Tradeline to inquiry conversion ratio |
| USEC\_TL\_ENQ\_RATIO | Unsecured Tradeline to inquiry conversion ratio |
| USEC\_TL\_ENQ\_RATIO\_24m | Unsecured Tradeline to inquiry conversion ration in last 24m |

**Liability Variables**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| NW\_ATM\_D\_AMT\_6m | ATM debit amount in last 6m |
| avg\_bal\_12m | Average Current Account & Savings Account Balance in last 12m |
| NW\_BRN\_C\_NO\_12m | Branch credit count in last 12m |
| NW\_BRN\_C\_NO\_6m | Branch credit count in last 6m |
| NW\_BRN\_C\_NO\_9m | Branch credit count in last 9m |
| NW\_BRN\_D\_AMT\_6m | Branch debit amount in last 6m |
| NW\_BRN\_D\_NO\_12m | Branch debit count in last 12m |
| Application\_vintage | Current Account & Savings Account Vintage |
| NW\_CASH\_C\_AMT\_6m | Cash credit amount in last 6m |
| NW\_CASH\_C\_NO\_6m | Cash credit count in last 6m |
| NW\_CHEQUE\_D\_AMT\_6m | Cheque debit amount in last 6m |
| NW\_CHEQUE\_D\_AMT\_9m | Cheque debit amount in last 9m |
| NW\_CHEQUE\_D\_NO\_12m | Cheque debit count in last 12m |
| NW\_CHEQUE\_D\_NO\_9m | Cheque debit count in last 9m |
| chq\_bounce\_count\_12m | Cheque\_BOUNCE\_COUNT\_12m |
| C\_variation\_12m | Co-efficient of variation of Current Account & Savings Account Balance in last 12m |
| C\_variation\_6m | Co-efficient of variation of Current Account & Savings Account Balance in last 6m |
| C\_variation\_9m | Co-efficient of variation of Current Account & Savings Account Balance in last 9m |
| Dummy\_application\_id | Dummy\_application\_id |
| min\_bal\_6m | Min Current Account & Savings Account Balancein last 6m |
| prn\_debit\_credit\_6m | proportion of Debit to Credit Amount in the last 6 months |
| prn\_debit\_credit\_9m | proportion of Debit to Credit Amount in the last 9 months |
| Prn\_NW\_ATM\_D\_AMT\_12m | Proportion of ATM debit Amount to total debit amount in the last 12 months |
| Prn\_NW\_BRN\_C\_AMT\_12m | Proportion of Branch Credit Amount to total Credit amount in the last 12 months |
| Prn\_NW\_CASH\_C\_AMT\_12m | Proportion of Cash Credit to Total Credit in th last 12 months |
| Prn\_NW\_CASH\_D\_AMT\_12m | Proportion of Cash Debit to Total Debit in th last 12 months |
| Prn\_NW\_CHEQUE\_C\_AMT\_12m | Proportion of Cheque Credit to Total Credit in th last 12 months |
| Prn\_NW\_DIGI\_D\_AMT\_12m | Proportion of Digital Credit to Total Credit in th last 12 months |
| Prn\_NW\_FT\_D\_AMT\_12m | Proportion of Fund Transfer to total debit in the last 12 months |
| Prn\_NW\_INB\_MB\_D\_AMT\_12m | proportion of Internet & Mobile debit count to total debit count in last 12 months |
| Prn\_NW\_NEFT\_IMPS\_C\_AMT\_12m | Proportion of NEFT and IMPS Credit to total Credit in the last 12 months |
| Prn\_NW\_NEFT\_IMPS\_D\_AMT\_12m | Prn\_NW\_NEFT\_IMPS\_D\_AMT\_12m |
| Prn\_NW\_ATM\_D\_NO\_12m | proportion of ATM debit count to total debit count in last 12m |
| Prn\_NW\_ATM\_D\_NO\_9m | proportion of ATM debit count to total debit count in last 9m |
| Prn\_NW\_BRN\_C\_AMT\_6m | proportion of Branch credit amt to total credit amt in last 6m |
| Prn\_NW\_BRN\_C\_No\_12m | proportion of Branch credit count to total credit count in last 12m |
| Prn\_NW\_BRN\_C\_No\_9m | proportion of Branch credit count to total credit count in last 9m |
| Prn\_NW\_BRN\_D\_AMT\_12m | proportion of Branch debit amt to total debit amt in last 12m |
| Prn\_NW\_BRN\_D\_AMT\_9m | proportion of Branch debit amt to total debit amt in last 9m |
| Prn\_NW\_CASH\_C\_No\_12m | proportion of Cash credit count to total credit count in last 12m |
| Prn\_NW\_CASH\_C\_No\_6m | proportion of Cash credit count to total credit count in last 6m |
| Prn\_NW\_CHEQUE\_D\_AMT\_9m | proportion of Cheque debit amt to total debit amt in last 9m |
| Prn\_NW\_DIGI\_D\_AMT\_6m | proportion of Digital debit amt to total debit amt in last 6m |
| Prn\_NW\_DIGI\_D\_NO\_9m | proportion of Digital debit count to total debit count in last 9m |
| Prn\_NW\_FT\_D\_NO\_6m | proportion of Fund transfer debit count to total debit count in last 6m |
| Prn\_NW\_FT\_D\_NO\_9m | proportion of Fund transfer debit amount to total debit count in last 9m |
| Prn\_NW\_INB\_MB\_D\_NO\_9m | proportion of Internet & Mobile debit count to total debit count in last 9m |
| Prn\_NW\_POS\_D\_AMT\_12m | proportion of POS debit amt to total debit amt in last 12m |
| Max\_Min\_Bal\_ratio\_9m | Ratio of Max to Min Current Account & Savings Account balance in last 9m |
| ever60\_24m\_StrictFlag | Target Definition |
| debits\_no\_overall\_12m | Total debit count in last 12m |
| total\_deposit\_no\_all | total\_deposit\_no\_all |

**Dev- development – train**

**OOT -out of time - Test**

**Demog Variables**

|  |  |
| --- | --- |
| **Variables** | **Description** |
| cnt\_coapplicant | Count of Co applicants |
| FOIR2 | Fixed Obligation to Income ratio |
| REQUESTED\_TENURE | Home Loans' Request Tenure |
| Doc\_form\_16 | If the applicants have submitted Form 16 or not |
| age\_max | In case of multiple applicants , maximum of their ages |
| income\_max | In case of multiple applicants , maximum of their monthly income |
| age\_min | In case of multiple applicants , minimum of their ages |
| income\_min | In case of multiple applicants , minimum of their monthly income |
| income\_sum | In case of multiple applicants , sum of their monthly income |
| edu\_max | In case of multiple applicants , the maximum of their education qualifications ( Higher the number , more the qualifications) |
| edu\_min | In case of multiple applicants , the minimum of their education qualifications ( Higher the number , more the qualifications) |
| Loan\_sub\_type\_1 | Loan Subtype |
| LTV | Loan to Value Ratio |
| Max\_dependent | Maximum of Dependents on Applicants |
| NATURE\_OF\_ORGANISATION | NATURE\_OF\_ORGANISATION |
| OCCUPATION\_TYPE | OCCUPATION\_TYPE |
| Organization\_type | Organization\_type |
| PROPERTY\_INSURANCE\_AMT | PROPERTY\_INSURANCE\_AMT |
| emi\_income\_sum | Ratio of EMI to Sum of Income , in case of multiple applicants we take sum of income |
| emi\_income\_max | Ratio of EMI to Income , incase of multiple applicants we take maximum of that ratio |
| SALARIED\_SELF | SALARIED\_SELF |
| SAL\_FLAG | Salary Flag |
| CHANNEL\_CODE | Sourcing Channel of the loan |
| City\_tier | Tier of the city , lower the better |
| ever60\_Flag\_Morat | Morat Adjusted Target Variable |
| AR\_Excl\_Flag | Status of Application |

**Business Reporting**

**Task 1.1 (SQL-Oracle)**

**Stage 1:**

1. Construct and ER-Diagram for the above-mentioned Requirement
2. Construct Tables as per the ER-Diagram.
3. Identify the relationships between tables and use appropriate standards for the same where applicable
4. Insert the appropriate data into the identified tables from the sample dataset provided.

**Task 1.2 (Informatica- ETL)**

Configure the source and target connections (Data Integration).

2. Implement Rank & Router transformations in single pipeline using ETL

3. Implement Filter & Aggregation transformations using ETL

4. Implement Flat file transferring to target using Aggregation

5.Create different sources & Targets integrate with ETL

**TASK 1.3(Visualization using Power-BI)**

**Connect the data with Power BI desktop and perform Data Manipulation using Power Query Editor. Perform the below tasks in Power BI Desktop.**

1. Prepare a report with Result in Power BI of given Question & Publish it into a Dashboard & Export Dashboard into a PPT.
2. Prepare a report by using appropriate Graph with Result in Power BI of given Question & Publish it into a Dashboard & Export Dashboard into a PPT

* RDD Analysis Results
* DF Analysis Results
* Hive Analysis Results
* Nosql Analysis Results

**Business Analysis**

**TASK 2.1 (Exploratory Data Analysis)**

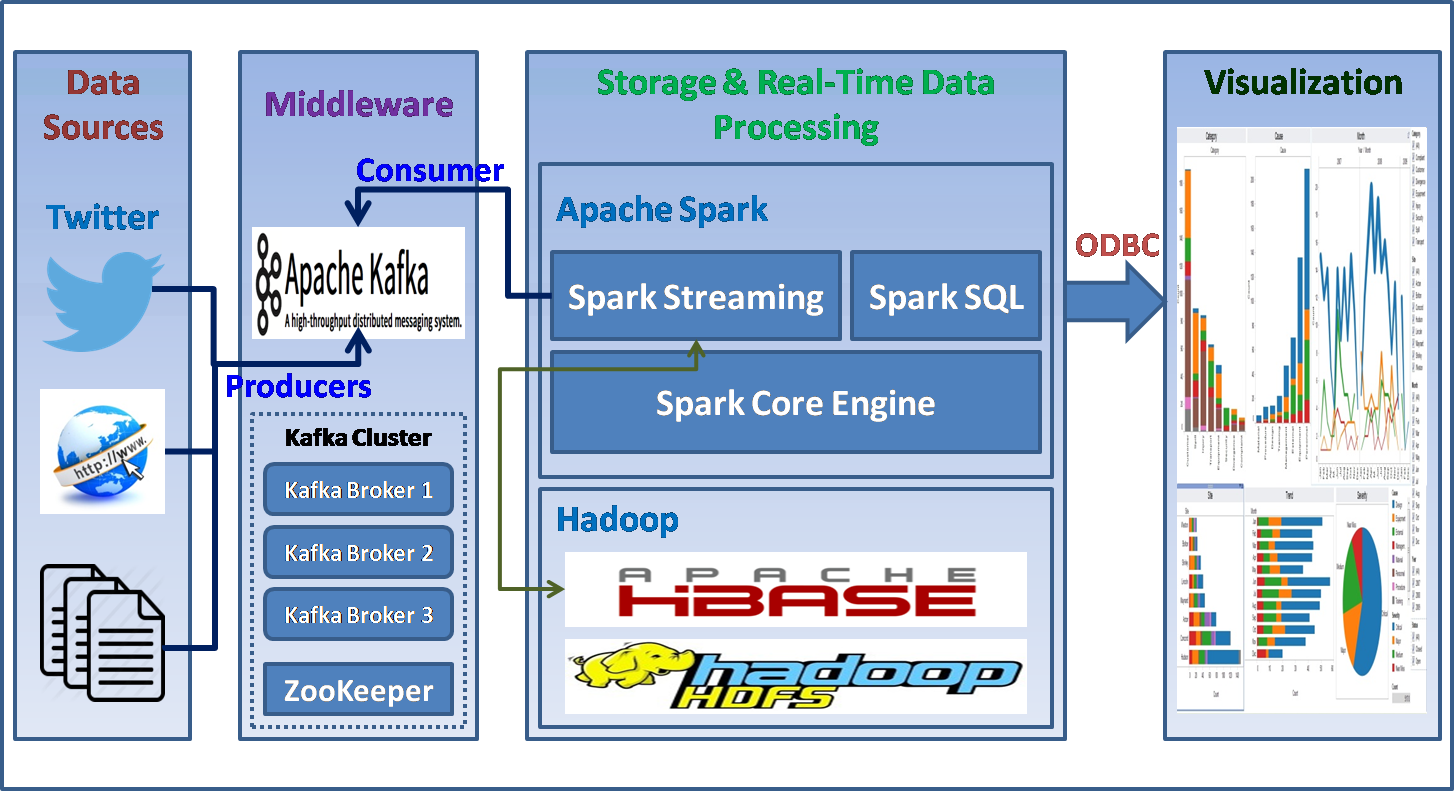
Data Preparation/Analysis tasks including (but not limited to) the following.

1. Univariate, Bi- Variate Analysis and Multi- Variate Analysis
2. Missing values identification and treatment
3. Outlier analysis and treatment
4. Data scaling using min-max and/or Z-score normalisation
5. Data transformation
6. Feature Engineering

**Task 2.2(Statistical Analysis)**

1. Descriptive statistics for both numerical and categorical and draw few insights from them.
2. Perform relevant hypothesis testing (t, chi-Square, Anova tests)

**Data Engineering**



**Task 3 - Data Analysis and Model building using Big Data Tools**

1. Big Data technologies like HDFS, Hive and PySpark need to be used as the historical data increases in size. As part of this task the following activities need to be done.
2. Develop a PySpark application to load data Spark DataFrames and save it into Hive tables on a Hadoop cluster in an optimized format.
3. Perform profiling of the data through PySpark and ensure that it is migrated correctly whereever the source is an RDBMS
4. Write PySpark routines to cleanse the data, prepare the data to handle missing values, and the data transformations identified in task 1.1 again making sure that the data is written into Hive tables in an efficient format
5. If the predictive model identified in task 4 available in Spark MLlib then develop a PySpark application to implement and evaluate the ML model identified with appropriate metrics\
6. Ensure that the best practices are followed and the design & code use the features of Spark and take advantage thereof.

**DataScience**

**Task 4(Model building using ML algorithms)**

**Model Building:**

1. Build an appropriate ML model/s on the data.
2. Compare various ML models with appropriate regularization and/or hyper-parameter tuning.
3. Evaluate the performance of the model.
4. Identify the right metric to evaluate the performance of the model.
5. Identify issues and concerns on the given data and suggest the best technique/s to overcome the issues.