**Capstone Project 1: Final Project**

**Proposal with problem statement:**

* What is the problem you want to solve?

The problem lies in getting the predicting the credibility of a person of getting a loan along with deciding the time-line when the loan would get paid off.

* Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn’t have done otherwise?

My client here is my company, who will be able to get the best lead who can be converted and given the best possible loan. Based on the analysis the company will be able to disburse the loan in the right manner with capability to decide the correct tenure and total sanction amount.

* What data are you using? How will you acquire the data?

The data is easily available from the company’s database. Will be referring to a collated data of the different attributes of a lead and then clean the data before processing. Since this is confidential data so changing the primary attributes (i.e.: email and phone number) with dummy data will act the input data set.

* Briefly outline how you’ll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.

As data wrangling becomes the first step, so i will start collecting the data and as mentioned the data is already available so probably referring to a single table in the database can fetch me the input data.

Now, the data captured, is from multiple sources so the entire data needs to be cleaned before use. We need to transform it with standard acceptable values and keep distinct and non duplicate data.

Once cleaning is completed will apply different algorithms to come up with a trend and use python libraries to present a visualization of the data.

* What are your deliverable? Typically, this includes code, a paper, or a slide deck.

**Deliverables would consist of:**

1. A working model that can take the input data set and generate patterns based on the historical data.
2. Report generated after applying different statistical Methods for Data Analysis
3. Also a machine learning trained algorithm which will be able to train the model based on existing/new data sets.

**Data collection and wrangling summary**

Getting the data was not a tuff job as it was readily available in the database. We were just supposed to clean the data and make minor changes.

Source:

Dumped the data from a financial DB to a csv file and took it as an input.

Name of the raw file: dataset\_uncleaned\_28072020.csv

Set of columns taken under consideration:

[ 'industry', 'disbursal\_date', 'tnc\_amount\_first', 'tnc\_ir\_first', 'tnc\_tenure\_first',

'min\_cibil\_score', 'loan\_type', 'd\_date\_of\_birth', 'vintage',

'3m\_avg\_bal', '3m\_avg\_bal\_new', 'total\_liab\_new', '90adb',

'l\_created\_date', 'tenure', 'score', 'rounded\_score']

Sample:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **industry** | **disbursal\_date** | **tnc\_amount\_first** | **tnc\_ir\_first** | **tnc\_tenure\_first** | **min\_cibil\_score** | **loan\_type** | **d\_date\_of\_birth** | **vintage** | **3m\_avg\_bal** | **3m\_avg\_bal\_new** | **total\_liab\_new** | **90adb** | **l\_created\_date** | **tenure** | **score** | **rounded\_score** | **disbursal\_date** | **getdate** | **mob** |
| **Building supplies and accessories** | 2018-05-11 05:30:00 | 180000 | 1.8 | 24 | 725 | ecom | 1991-06-01 | 12 | 87212.7466666667 | 88753.76 | 0 | 88753.76 | 2018-05-03 05:42:12.601 | 19 | 1.47368421052632 | 1 | 2018-05-11 05:30:00 | 2020-07-28 16:49:28 | 1.10681700517722 |
| **Industrial Supplies & Solutions** | 2018-07-09 05:30:00 | 700000 | 1.8 | 24 | 644 | ecom | 1973-03-20 | 37 | 89512.4 | 87169.43 | 0 | 87169.43 | 2018-06-27 05:25:36.939 | 17 | 1.29411764705882 | 1 | 2018-07-09 05:30:00 | 2020-07-28 16:49:28 | 1.0261718438869 |
| **Computers,Mobile & related Accessories** | 2018-09-05 05:30:00 | 270000 | 2 | 24 | 741 | ecom | 1987-11-08 | 55 | 40480.3533333333 | 43564.94 | 350000 | 43564.94 | 2018-08-22 08:27:54.742 | 15 | 1.4 | 1 | 2018-09-05 05:30:00 | 2020-07-28 16:49:28 | 0.948214854639586 |
| **Other Service Activities** | 2018-07-27 05:30:00 | 50000 | 1.8 | 12 | 649 | ecom | 1983-09-26 | 12 | 12079.0066666667 | 11838.57 | 264313 | 11838.57 | 2018-06-04 05:13:18.035 | 12 | 1.25 | 1 | 2018-07-27 05:30:00 | 2020-07-28 16:49:28 | 2.0039565909996 |
| **Home, Furnishing and Decor** | 2018-07-24 05:30:00 | 70000 | 2.1 | 12 | 694 | ecom | 1987-12-31 | 22 | 8215.35333333333 | 8030.16 | 64959 | 8030.16 | 2018-06-13 10:30:35.362 | 12 | 1.41666666666667 | 1 | 2018-07-24 05:30:00 | 2020-07-28 16:49:28 | 2.01202110712863 |

Cleaning steps:

* **Inserted mean values in place of null/nan values**
* **Converted date columns to type datetime**
* **Chose numerical columns only for analysis (excluding categorical)**
* **new\_columns = ['disbursal\_date', 'tnc\_amount\_first',**

**'tnc\_ir\_first', 'tnc\_tenure\_first', 'min\_cibil\_score',**

**'d\_date\_of\_birth', 'vintage', '3m\_avg\_bal', '3m\_avg\_bal\_new',**

**'total\_liab\_new', '90adb', 'l\_created\_date', 'tenure']**

* **Normalized data via min-max**

Finally the cleaned data was saved to **‘cleaned\_dataset.csv’.**

**Exploratory data analysis summary:**

Univariate Analysis:

* Univariate analysis includes exploration of data wrt a single variable.
* This includes histograms, pie charts, bar charts, box plots.
* Histograms: continuous data displayed as frequency charts based on intervals
* Barcharts: Categorical data within a variable can be displayed in multiple columns
* Pie charts: frequency distribution of multiple categories as a circle. (Use only when less categories are present)

Boxplot: Representation displaying the medians, quartiles, and upper/lower bounds for outliers.

[When Should You Delete Outliers from a Data Set? - Atlan | Humans of Data](https://humansofdata.atlan.com/2018/03/when-delete-outliers-dataset/)

Bivariate analysis:

* Correlation tests with Covariance.
* Scatter plot analysis. (Has to be checked. There might be multiple segments of data each providing a different analysis). If there is enough support for the outliers/multiple segments, each of them might require a different analysis.

**min\_cibil\_score**: from the data 7.3 % of the population have min cibil score -1.

This cibil score is given to users who don't have any credit history at all. By trying to fit people with non existent credit history in the same model might create errors. Thus a new subset for this has to be created for people whose credit score is -1 and whose isnt. This population of 7.3% may influence the balance of the model. The min\_cibil\_score follows a bell curve(normal distribution) for the values that are not equal to -1. Hence this distribution can have a model of it’s own.

A screenshot of a cell phone

Description automatically generated

On checking the distribution of data by splitting on the basis of cibil score equal to -1 or not, We find no influence/correlation with other variable. Hence min\_cibil\_score for applications=-1 can also be resolved by replacing it with median or statistically proven quantile for financial risk analysis.

**d\_date\_of\_birth**:

The distribution is skewed showing a majority of users born between 1978 and 1995 using the service indicating users between the age of 23 and 40 i.e gen y and very few gen x.

However this graph also indicates the working population distribution by age thus not providing a lot of insights.

**A close up of a logo

Description automatically generated**

**tnc\_tenure\_first**: The distribution of tenures:

A screenshot of a social media post

Description automatically generated

Thus with such an imbalance of tenures, this can be another attribute to be taken to create different models for each individual tenures of 24 and 12 months and different for another.

**Correlation factors:**

The amount and interest rates have negative correlation. This implies the higher the loan amount the lower the interest rate applicable on it.

The tenure and amount have positive correlation. This trend implies that higher the tenure the loan amount would also be higher.

Correlation is also seen between 3m\_avg\_bal\_new and amount. Implies higher loan amounts are disbursed only to people with higher average recently recorded balance.

On comparing based on loan types.

The interest rates for lvl based loans are higher than ecom.

Yellow is loan\_type=‘lvl\_loan’ and blue is loan\_type=‘ecom’

A close up of a logo

Description automatically generated

**Boxplot analysis:**

Plots show insignificant insights regarding whether a variable is an outlier or not. Since these are correlated dependent variables replacing the particular field with mean/median is not enough. The correlated variables also have to be changed. Thus in this case the record has to be dropped. But since the data is required/mandatory to prevent overfitting of data. Specific case data cannot be excluded. If there were more data points for these outliers to have their own model, data can be segmented for certain intervals

**Results and In-depth analysis using machine learning**

From the categorical scoring based models different types of models were trained and the confusion matrices were analysed.

    1   2  3  4  5  6

1  [27  2  2  0  1  3]

2  [ 6 21  4  1  1  2]

3  [ 1  1 31  0  5 11]

4  [ 3  2  4 25  5  6]

5  [ 0  0  2  1 42 11]

6  [ 0  0  3  0  5 62]

1. The above confusion matrix tells how many of which category were predicted as expected.
2. The diagonal elements imply that the result and expected are the same.
3. The rows are the expected values and columns are the resultant values.
4. This is a risk based model with a higher score implying lower risk.
5. The upper triangular matrix indicates that a model is classifying lower rated users as higher and is thus not acceptable. The lower triangular matrix indicates that a model is classifying higher rated users as lower. This implies losing on business. Since money lending is a risk evaluation based sector, losing possible customers is better than defaulters hence the aim is to improve the diagonal matrix and minimize the upper matrix as much as possible.

**KNNClassifier** gave best results with n =1, 1 closest neighbor was good for data rather than multiple. This tells us that the data is highly continuous and clusters overlap via this method and may result in incorrect results when considered multiple results.

**Decision Trees** performed poorly in training and testing thus resulting in a model with high variance and bias.

**RandomForestClassifier** with it’s ensemble like procedure improved upon these weaker points of Decision Trees by aggregating the results in the last point.

**On comparing Ensemble of n decision trees and Random Forest of n estimators**: The Random forest gave better results thus implying the nodes right before the leaves have higher influence and should be optimized to give better results than generating completely new random trees itself.

**MLP classifier** with different Hyperparameter tuning via grid search resulted in highly unfeasible training time for the dataset in hand still returning unsatisfactory results.

Summary:

* Due to irregular distributions of data across the dataset for each variable.
* Some following unimodal, some bimodal, some uniform normal, some skewed normal. Naive bayes classifiers performed poorly. But by using the prior probabilities and tuning for Complement/Multinomial Naive Bayes it can be used in ensemble voting models to improve in the weaker segments of other models.
* Gradient boosting classifier was on par with the RandomForest Classifier.
* Implementing a voting classifier to improve the bias of both classifiers resulted in little improvement .

**How to improve (Future scope):**

Further improve the analysis of variables by using transformations(PCA, LDA, Autoencoders). Models based on specific segments of data to prevent generalization/underfitting for the whole of the dataset.

Experimenting with ensemble of linear and nonlinear classifiers to improve accuracy.

Better encoding of timestamp fields extracting useful information based on transformations.

Utilizing categorical data by encoding it into numerical (binary/ onehot/ label)