Decision System for Credit Underwriting

**Presented by:**

### Raja N. (Team Lead)

### Snehal Tambe (Project Coordinator)

### Sai Harsha

### Mandar Vast

### Jagdish Kharatmol

### Parth Sethi

**Abstract:**

People applying for loans have increased significantly in the last few years and we need to have a system which can validate the client’s credit risk. Risk here is a combination of customer worth, behaviour, and intention.

Using the data provided, we came up with a scientifically developed data driven system to automate decisions and processing for stronger and quicker underwriting. The process helped in gaining more insights from the data that help in improving various metrics including marketing, risk, product, and customer experience.

Our Underwriting system takes into consideration various dimensions such as financial, demographic and behavioural aspects of customers for risk assessment.

### 

### 

### 

### 

### 

### 

### 

### 

### Table of content

[What is Credit Underwriting?](#_heading=h.gmsk9l9m5q9y) 5

[Background and Motivation](#_heading=h.30j0zll) 5

[Problem statement](#_heading=h.1fob9te) 6

[Age Distribution](#_heading=h.tyjcwt) 6

[Defaulter income type](#_heading=h.efac96qjtab9) 7

[Education Type of the Clients](#_heading=h.hzv9flb199xv) 8

[Organization Type](#_heading=h.jf1655db9gkx) 8

[Top Occupation Type Defaulters](#_heading=h.2s8eyo1) 9

[Organization Type w.r.t Occupation Type](#_heading=h.fbwphg8sqly2) 9

[Education Type w.r.t Income Type](#_heading=h.3rdcrjn) 10

[Project Architecture](#_heading=h.bhp0fcet2nlb) 12

[Feature Engineering](#_heading=h.lnxbz9) 12

[Data Summary](#_heading=h.bhawi5669ko1) 13

[Data Pre-processing steps](#_heading=h.1ksv4uv) 13

[Data Merging](#_heading=h.44sinio) 14

[Handling Null Values](#_heading=h.2jxsxqh) 14

[Feature Encoding](#_heading=h.q3pvk3dx4dh9) 14

[Transforming Days Columns](#_heading=h.v0cm7t7y8g9r) 15

[Standardization](#_heading=h.1y810tw) 15

[Model building](#_heading=h.4i7ojhp) 15

[Different performance measure metrics used](#_heading=h.1ci93xb) 16

[Cross Validation](#_heading=h.3whwml4) 16

[KS Statistics for different models](#_heading=h.m0c2af3pq8qh) 16

[Logistic Regression](#_heading=h.qsh70q) 16

[Gradient Boost](#_heading=h.hc7i0zyxh37m) 17

[Xtreme Gradient Boost](#_heading=h.8ui17bdezr2p) 19

[User interface](#_heading=h.vy6cuigdv1y3) 20

[Business Insights](#_heading=h.x0ord0qycd0y) 21

Future scope 22

Reference 23

### 

### 

### 

### 

### 

### 

### 

### What is Credit Underwriting?

Underwriting is the process by which the lender decides whether an applicant is creditworthy and should receive a loan. An effective underwriting and loan approval process is a key predecessor to favourable portfolio quality, and a main task of the function is to avoid as many undue risks as possible.

The underwriting system is like a financial expert who takes a look at your finances and assesses and specifies how much risk a lender will take on if they decide to give you a loan.

When we have a large number of applications the traditional approaches would take time and the client won’t wait if he requires a loan urgently and he may approach other places which will be a loss for the company. The contemporary system had evolved with the advent of technology and more data.

### Background and Motivation

Numerous banks revamped their credit underwriting process with a focus on speed, costs, efficiency, and customer satisfaction but they still face the risk of a customer not paying back the credited amount. Therefore, there is a need to develop a tool which can assist the banks in choosing the right customers. The above-mentioned problem can be addressed with the proposed system, which can be tuned further for more robust underwriting. This project provides the overall idea of Data Analysis, Machine Learning, and its application in credit business.

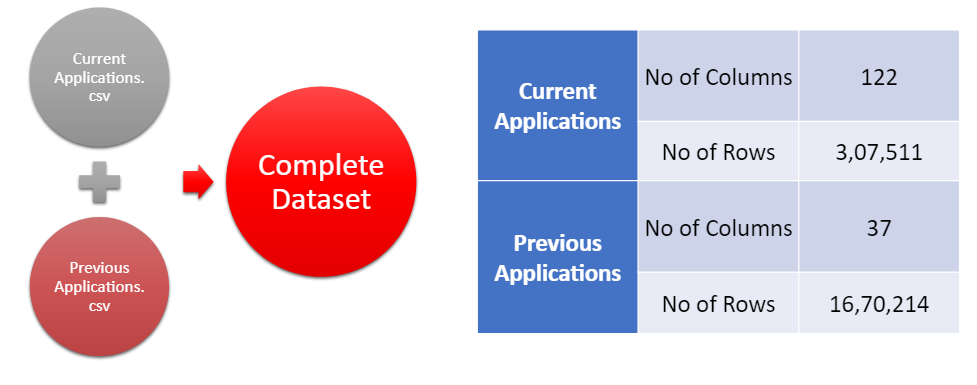
### Problem statement

Credit Karma seeks to evaluate their potential customers for risk to grow their lending portfolio. Strong under-writing is very important in reducing risk and growing their portfolio which is why a detailed decision is needed for the business. What is currently being written are manual and standard rules that are professionally set, time-consuming, general and prone to human error and bias. They want to establish a system that is scientifically automated to make decisions and to process solid and fast writing.

**Objectives**

* To understand and explore the data
* Dive deeper into the data for hidden patterns. To develop a machine learning model that will lower the risk for the credit lenders and provide a robust underwriting system
* To drive business insights that will help the management to make better decisions.

**Dataset details**



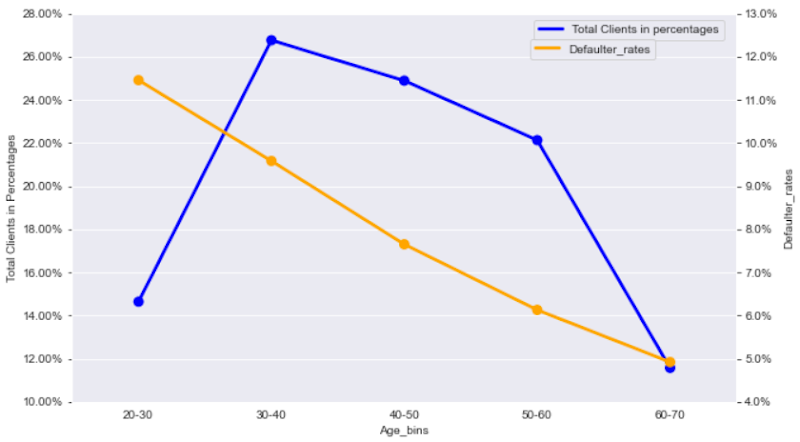
* Our Datasets consisted of Client’s **Financial data**, **Credit Scores** from 3 different Credit Bureaus, Flags for various **Contact info**, Flags for **Documents submitted**, applicant’s **Property info, etc**.
* We didn’t have any of the **Transaction data**.

**Dataset’s Source** : <https://www.kaggle.com/c/home-credit-default-risk/data>

**EDA**

* In the Exploratory Data Analysis, we performed an **initial investigation** on data so as to discover patterns, to spot anomalies, to test hypotheses and to check assumptions with the help of summary statistics and graphical representations.
* Some Important insights we have gained from this step are:

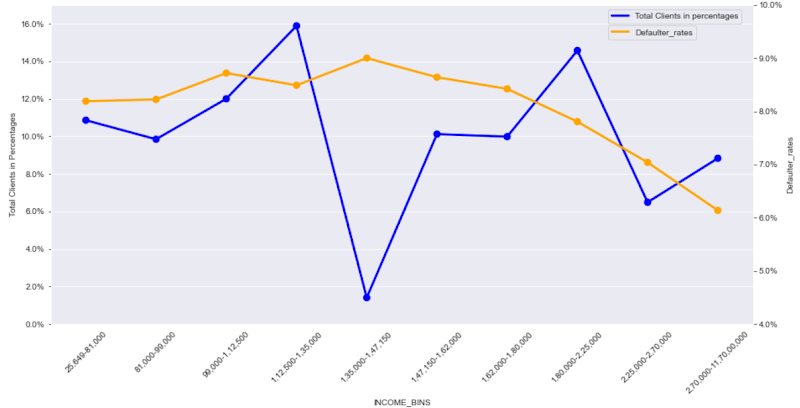
### Age Distribution



* The age is inversely corelated to the risk, i.e., higher the customers age, lower the risk.
* Majority (~74%) of the customers are between 30 years and 60 years.
* About 27% of our clients are aged between 30-40 years and have a risk of **9.5%**.
* The 40-50 age group has a very low default rate of about **7.6%.**
* Similarly, clients aged 50-60 have a high number of clients but the lowest default rate is **6%.**

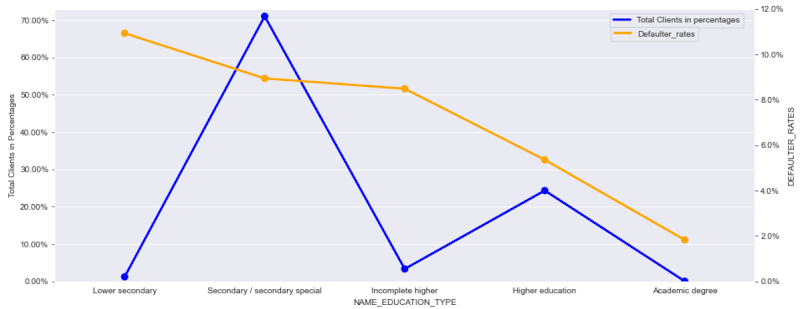
### 

* **Defaulter income type**



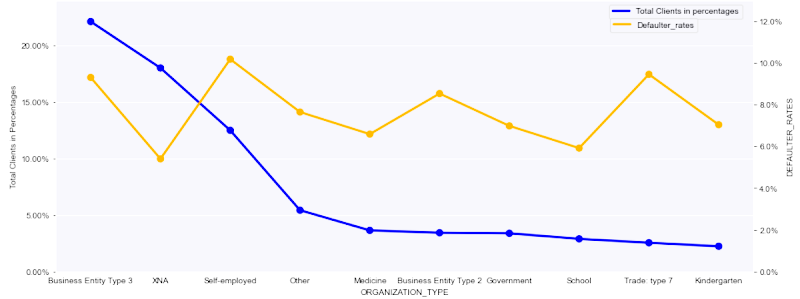
* **16%** of our customers have an income between 1,12,000-1,35,000 with a high default rate of **8.5%**.
* While, we have relatively low client numbers with a revenue between 1,80,000-2,25,000 but with a relatively lower default rate of **7.8%.**

### Education Type of the Clients



* Even though we have a lower number of clients from Incomplete, higher education at **3.3%**, they have a high default rate of **8.5%**.
* **71%** of our clients are Secondary/ Secondary special graduates with a higher risk of **8.9%.**
* Higher education clients make huge profits even though we have a large number of clients with highereducation compared to those with Incomplete higher education, their default level is much lower.

### Organization Type

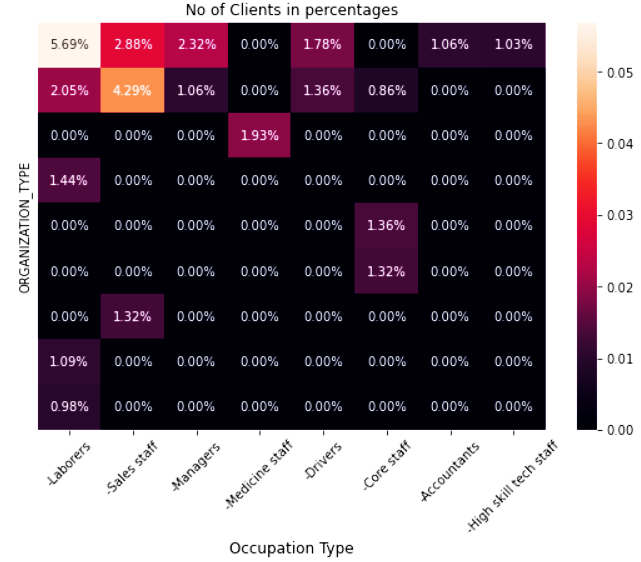


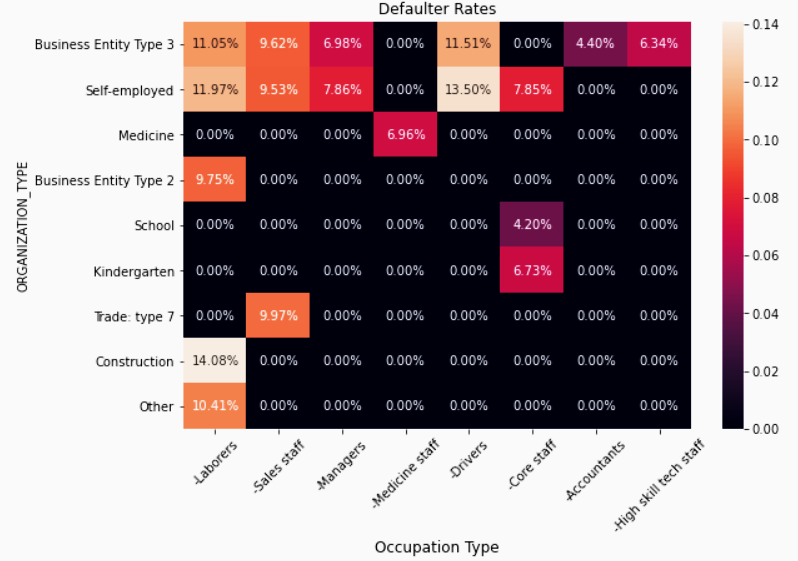
* 22% of our clients are from Business Entity Type 3 with a high level of error of **9.3%.**
* XNA has slightly lower customer numbers but the number of subscribers is much lower compared to **5.4%.**
* There is a very low number of self-employed customers but with a very high automation rate at **10.1%**

### Top Occupation Type Defaulters

* Around 18% of our clients are Laborers with a risk of 10.6%.

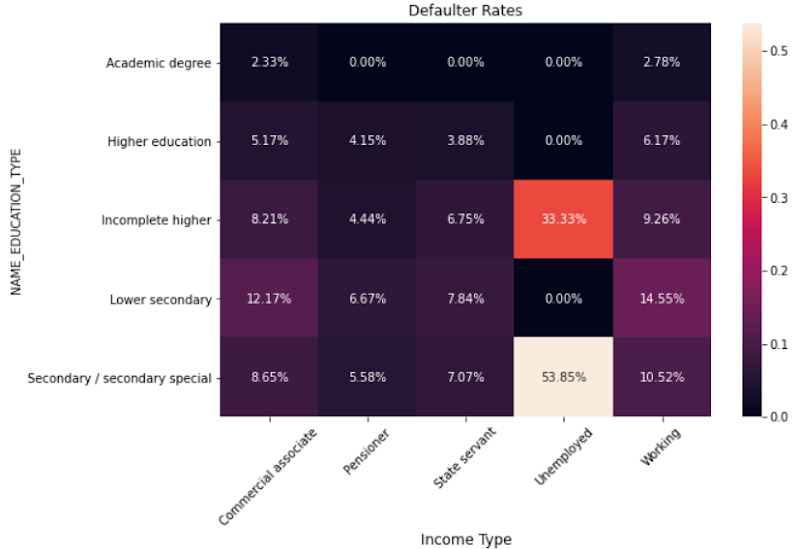
### Organization Type w.r.t Occupation Type

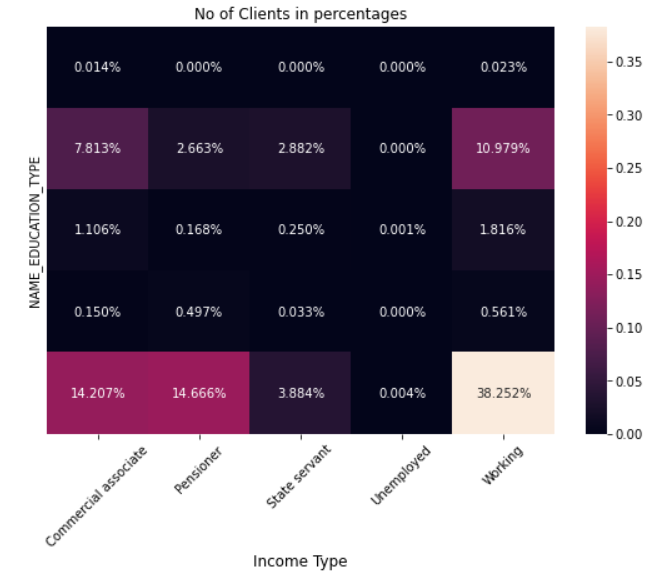




* 5.7% of our clients are employees from Business Entity type 3 and have a very high rate of defaulter rate at **11%**.
* Self-employed or Business Entity type 3 Employees are also ranked with the highest numbers (4.3% and 2.9%) and have the highest defaulter rate of **9.5%** and **9.6%** respectively.
* So we have to focus a little bit on these types of customers and change our collection strategies accordingly.

### Education Type w.r.t Income Type





* 38% of our clients are from the Working class who are Secondary/ Secondary special graduates but they have a high risk of **10.5%**.
* We have a slightly lower number of clients of the Pensioner class who are Secondary/ Secondary special graduates at 14.6% but their default rate is very low at just **5.58%.**
* Similarly for working clients with higher education and Commercial Associate and Secondary Education, with a lower level of automation compared to Working class and Secondary Education.

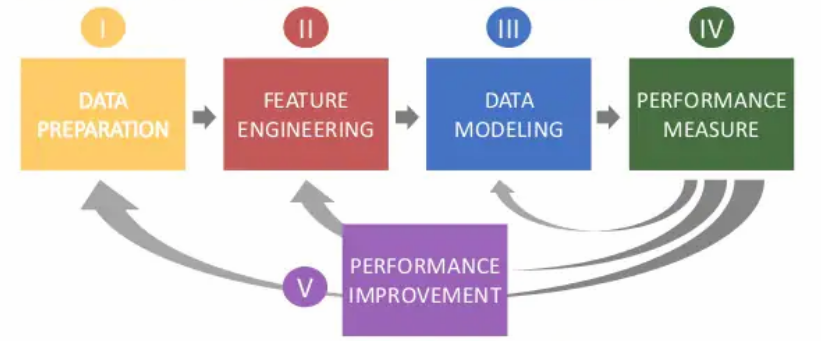
### Project Architecture

Feature Engineering

Data Preprocessing

Data Modeling

Model Evaluation



Performance   
Improvement

### Feature Engineering

Feature Engineering is one of the most important processes after data purification and integration. As in this process, we discover new features using existing ones. It helps in model prediction.

Key features based on feature engineering are:

**Number of Previous Applications**: No of previous applications of the current applicants.

**Credit to Income Ratio:** Ratio of Credit amount to Income amount.

**Annuity to Income Ratio:** Ratio of Annuity amount to Income amount.

**Goods Price to Income:** Ratio of Goods Price to Income amount.

**No of Documents Submitted:** Total No of Documents submitted out of the total 20 documents.

**Source Scores Average:** Average of External Score 1, External Score 2, External Score 3**.**

**Asset Index:** Index of the assets owned by the applicant (Income, Property, Car).

**Contact Info:** Total number of contacts provided by applicants (Mobile number, Work phone number, home phone number, Email).

**K means clusters:** Segregated data in 3 clusters based on elbow method and added their cluster number as a new column.

### Data Summary

* We have two sets of data; one **Current Applications** and other **Previous Applications**.
* The current application database has **3,07,511 rows** and **122 columns** including columns such as id, income, debt, pension rates, various contact information flags and documentary flags, columns related to region and city of the applicant, credit points, organizations and types of activity, and and multiple columns related to the applicant's location details.
* While the database for previous applications has **16,70,214 rows** and **37 columns.**

### Data Pre-processing steps

Data Merging

Handling Null Values

Feature Encoding

Standardization

Modelling

### Data Merging

* The process of merging two or more data sets into one dataset. We had raw data stored in several Comma-divided files and in order to obtain important data and a machine learning model, the team decided to merge the two databases. In addition, the purification of integrated data can be used more effectively.
* There were two sets of data for this project, the first being the Current Application Dataset and the other was the Previous Application Dataset. The Current Request dataset contains information about the current client application and the Previous Application database contains customer information about past loan applications. We have combined these two sets using Inner Join.

### 

### Handling Null Values

* Most of the machine learning model we want to use will give an error when we transfer Nan values ​​to it.

In this step, we have decided too:

* Drop columns with values ​​greater than **47%** of empty values.
* In columns with null values ​​between **5.35%** and **47%,** it means imputation is applied to continuous variable and mode input using the variable variable.
* Reduced column lines with null values ​​of less than **5.35%.**

### Feature Encoding

* Classes reduced to columns with more than 10 sections to 10 classes.
* Recording columns labeled with **One Hot Encoding** and ordinal columns with **Label Encoding**.

### 

### Transforming Days Columns

* These columns had days with negative numbers. We've taken absolute measures these days.

### Standardization

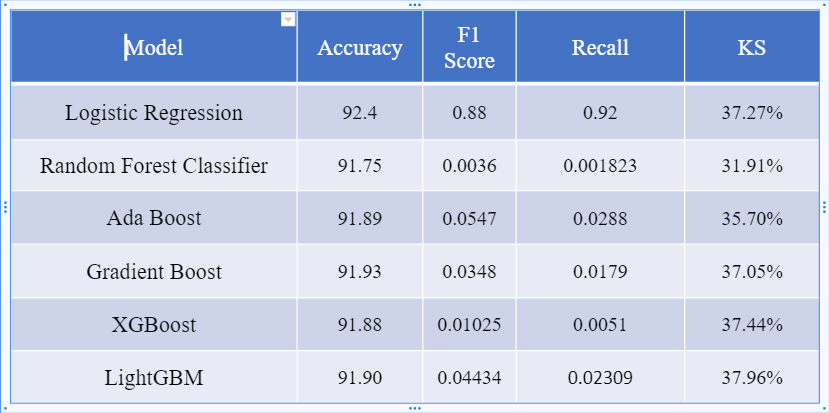
* Standardization is a scaling technique where the values are centered around the mean with a unit standard deviation.

Taking down all elements in the same scale i.e a value between 0 and 1.

### Model building

* At this stage the team has developed data sets for training, evaluation and goal setting. These data sets allow us to improve the analytical approach and train it, while setting aside other data for model testing.

**Different models with accuracy**

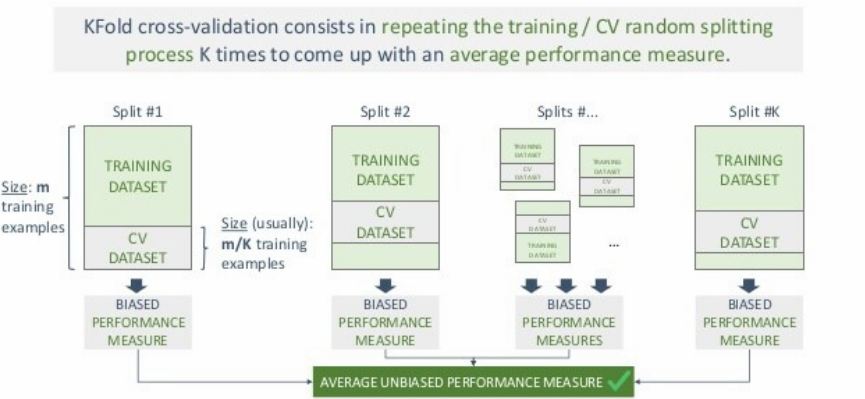


### Different performance measure metrics used

1. Roc Curve
2. F1 Score
3. Confusion Metrics
4. Precision
5. Accuracy

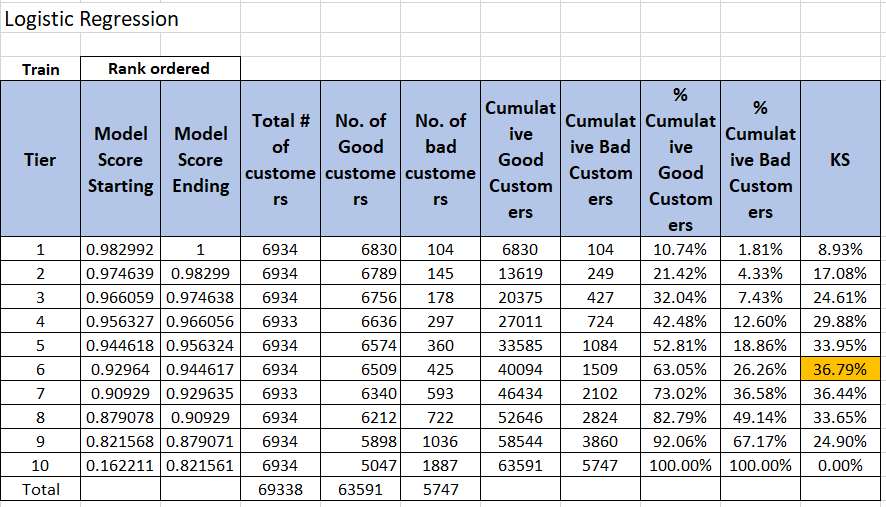
### Cross Validation

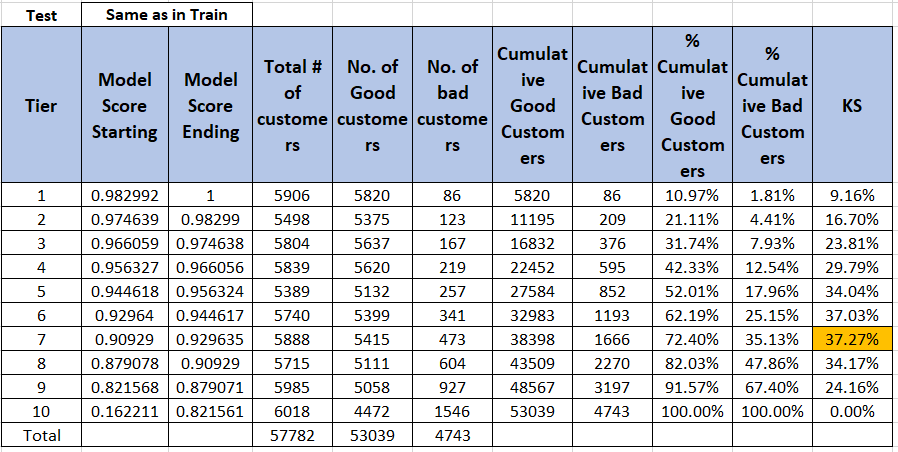
* To solve the problem of overfitting we use cross-verification. For validation, you create a certain number of folds, running a model in each fold.

****

### 

### KS Statistics for different models

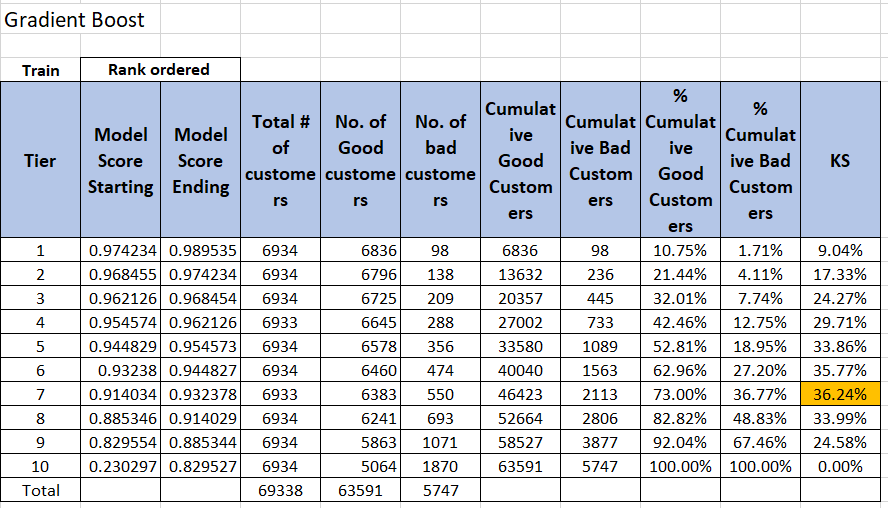
**Logistic Regression**

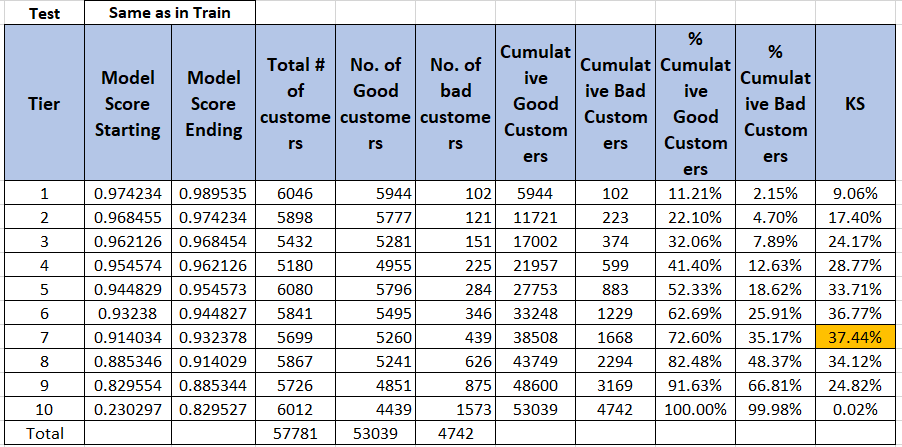
****

### 

### 

### 

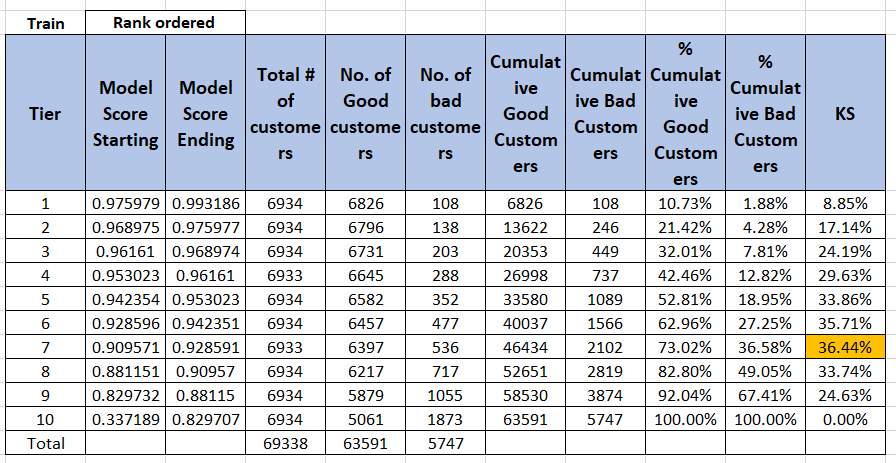
**Gradient Boost**

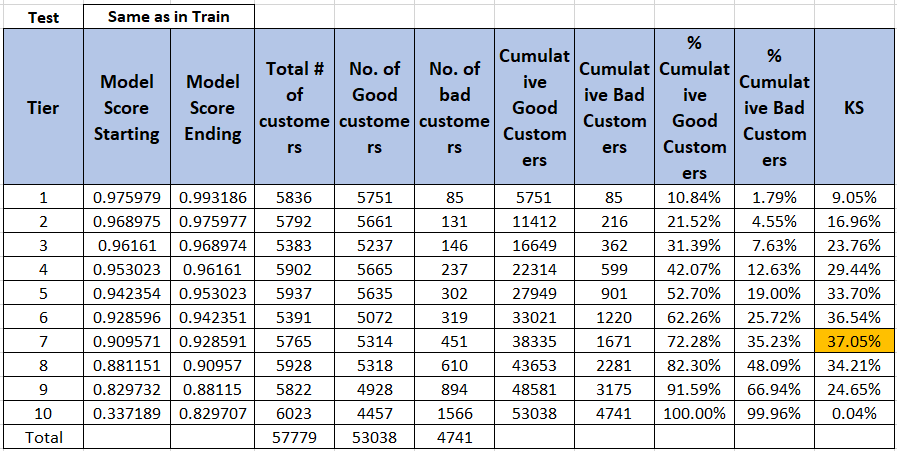


### 

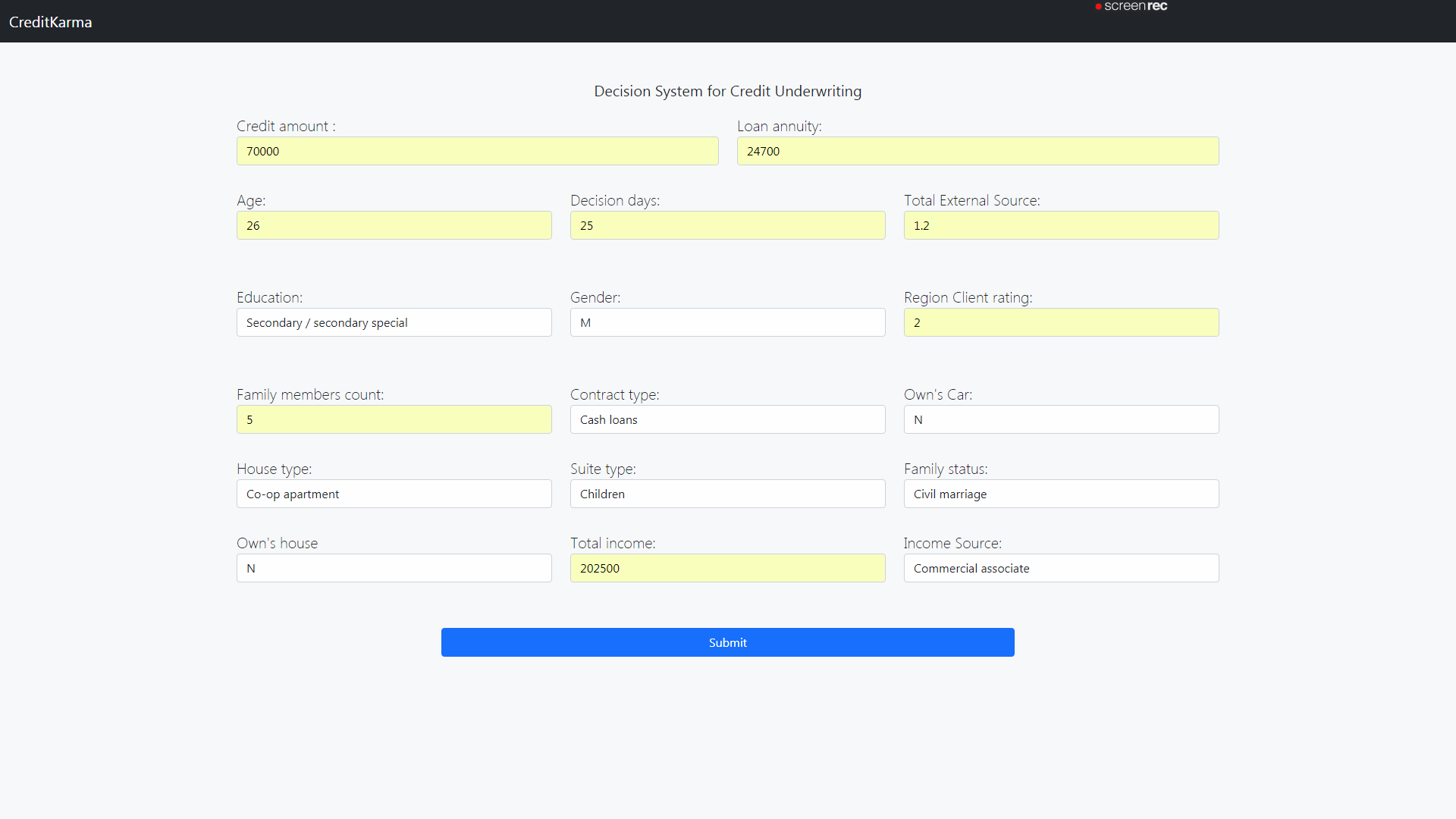
### 

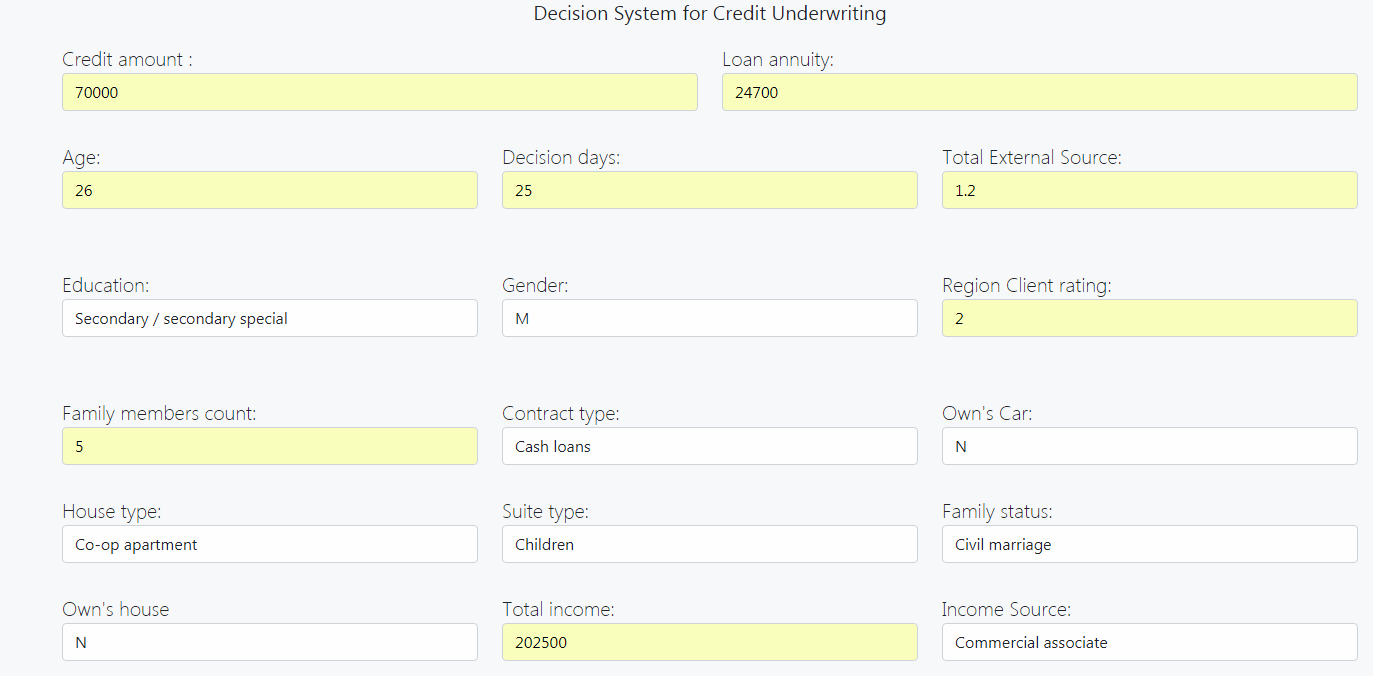
### Xtreme Gradient Boost





### User interface





### Business Insights

|  |  |
| --- | --- |
| **Finding** | **Actions need to be taken.** |
| As the age increases the default rate goes down. | We should target clients with ages between 40 to 70 years. |
| “Region rating of client” is working well. Which properly classifies the client good and bad client. | We can target the client with ratings as 1 and 2. As they have a higher proportion of data but lower default. |
| Education plays an important role in repayment of loans. | Priority of client based on education   1. Secondary/secondary special 2. Higher Education 3. Incomplete higher |
| The rejection for repeater is majorly due to “HC” code and “limit” code and for approval we have “XAP” code. | Can validate reasons for rejection for better results. |
| External score is highly correlated with default rate. | Sometimes decisions for loans need to be made in seconds. We can rely on External scores. |
| The client with a loan amount between 33000 to 34000 will cancel the loan application. | Don’t waste resources on those clients as they will cancel the loan. |

### 

### 

### Future scope

1. **Data Analysis** and **Feature Engineering** should be done in depth and include more features for hypothesis.
2. Can try to solve the problem with a **deep neural net** model which will surely increase the performance.
3. We can try to collect more data like the applicant’s **transaction data** which will further improve our model’s performance.

**Reference websites**

1.Stackoverflow

* Stack overflow was majorly referred for solving errors.

2.TowardsDataScience

* [**https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114**](https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114)
* [**https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15**](https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15)

3. listendata

* https://www.listendata.com/2019/07/KS-Statistics-Python.html