Project Presentation

By:- Shashank Kumar,

Shreyash Adhikari,

Mohammad Arafat Khan,

Siddhartha Ghosh

Semester – 6th

BE Computer Science

BMS College of Engineering, Bangalore

Customer Segmentation and Customer Lifetime Value in Banking Sector

Banks realize the need to change from a **product-centric** approach to **customer centric**.

Customer segments enable you to understand the patterns that differentiate customers

Types of Customer Segmentation criteria:

- Customer value
- Demographics
- Life stage
- Attitude
- Behaviour

Customer Lifetime Value – (CLV)

- Italian economist **Vilfredo Pareto** states that 80% of the effect comes from 20% of the causes, this is known as 80/20 rule or Pareto principle. Similarly, 80% of company's business comes from 20% customers.

Importance of CLV

- Companies need to identify those top customers and maintain the relationship with them to ensure continuous revenue. In order to maintain a long-term relationship with customers, companies need to schedule loyalty schemes such as the discount, offers, coupons, bonus point, and gifts.
- It is an aspect of Targeted Marketing.

Basic Formula

- *Gross margin * (Retention rate / [1+ Rate of discount – Retention rate]*

- *Gross Margin: Total revenue minus cost of acquisition and retention*
- *Retention Rate: Ratio of number of retained customers to number at risk*
- *Rate of discount: Interest rate used to calculate the present value of the future cash flow*

Key Inputs for Calculations

- Average balances of loans and savings on a per customer basis
- Average interest rate margin (%)
- Average income/revenue per customer generated from non-interest income sources (e.g. fees, commissions, and other sales)
- Costs of providing customer services and access (which include transaction costs, statement costs, infrastructure costs etc.)

Problem Statement

- Building various Machine Learning Models that will be able to predict CLV by learning from the previous data. And choosing the most appropriate model, based on its accuracy score.

Requirement Specification

Hardware Requirements:

- Processor : Intel i5
- Hard Disk : 120 GB
- Input Devices: Keyboard, Mouse , Monitor

Software Requirements:

- Operating system: Windows 10
- Tools: Jupyter Notebook, Anaconda Distribution
- Language: python 3
- Libraries: **sklearn**, **numpy**, **matplotlib**, **pandas**, **keras**

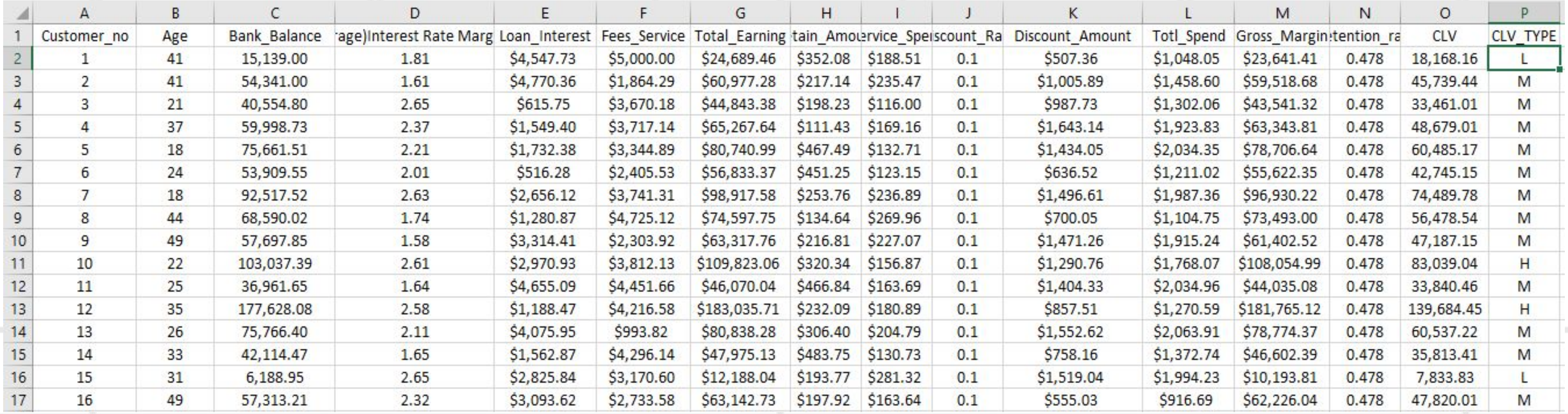
Objectives ‼

- Obtaining Dataset
- Identify the Problem Type
- Brief analysis on the Models to be used
- Create a High Level Design, Use Case Design,
- **Building a Machine Learning Model That will be able to predict CLV by learning from the previous data.**
- **Analysing the Accuracy**
- Visualization of the Dataset
- Data Set Analysis

Data Acquirement

- Kaggle
- GitHub
- DataCamp
- **Mockaroo***

Snapshot of the Dataset



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Customer_no	Age	Bank_Balance	age)Interest Rate Marg	Loan_Interest	Fees_Service	Total_Earning	tain_Amou	rvice_Spe	scount_Ra	Discount_Amount	Totl_Spend	Gross_Margin	tention_ra	CLV	CLV_TYPE
2	1	41	15,139.00	1.81	$4,547.73	$5,000.00	$24,689.46	$352.08	$188.51	0.1	$507.36	$1,048.05	$23,641.41	0.478	18,168.16	L
3	2	41	54,341.00	1.61	$4,770.36	$1,864.29	$60,977.28	$217.14	$235.47	0.1	$1,005.89	$1,458.60	$59,518.68	0.478	45,739.44	M
4	3	21	40,554.80	2.65	$615.75	$3,670.18	$44,843.38	$198.23	$116.00	0.1	$987.73	$1,302.06	$43,541.32	0.478	33,461.01	M
5	4	37	59,998.73	2.37	$1,549.40	$3,717.14	$65,267.64	$111.43	$169.16	0.1	$1,643.14	$1,923.83	$63,343.81	0.478	48,679.01	M
6	5	18	75,661.51	2.21	$1,732.38	$3,344.89	$80,740.99	$467.49	$132.71	0.1	$1,434.05	$2,034.35	$78,706.64	0.478	60,485.17	M
7	6	24	53,909.55	2.01	$516.28	$2,405.53	$56,833.37	$451.25	$123.15	0.1	$636.52	$1,211.02	$55,622.35	0.478	42,745.15	M
8	7	18	92,517.52	2.63	$2,656.12	$3,741.31	$98,917.58	$253.76	$236.89	0.1	$1,496.61	$1,987.36	$96,930.22	0.478	74,489.78	M
9	8	44	68,590.02	1.74	$1,280.87	$4,725.12	$74,597.75	$134.64	$269.96	0.1	$700.05	$1,104.75	$73,493.00	0.478	56,478.54	M
10	9	49	57,697.85	1.58	$3,314.41	$2,303.92	$63,317.76	$216.81	$227.07	0.1	$1,471.26	$1,915.24	$61,402.52	0.478	47,187.15	M
11	10	22	103,037.39	2.61	$2,970.93	$3,812.13	$109,823.06	$320.34	$156.87	0.1	$1,290.76	$1,768.07	$108,054.99	0.478	83,039.04	H
12	11	25	36,961.65	1.64	$4,655.09	$4,451.66	$46,070.04	$466.84	$163.69	0.1	$1,404.33	$2,034.96	$44,035.08	0.478	33,840.46	M
13	12	35	177,628.08	2.58	$1,188.47	$4,216.58	$183,035.71	$232.09	$180.89	0.1	$857.51	$1,270.59	$181,765.12	0.478	139,684.45	H
14	13	26	75,766.40	2.11	$4,075.95	$993.82	$80,838.28	$306.40	$204.79	0.1	$1,552.62	$2,063.91	$78,774.37	0.478	60,537.22	M
15	14	33	42,114.47	1.65	$1,562.87	$4,296.14	$47,975.13	$483.75	$130.73	0.1	$758.16	$1,372.74	$46,602.39	0.478	35,813.41	M
16	15	31	6,188.95	2.65	$2,825.84	$3,170.60	$12,188.04	$193.77	$281.32	0.1	$1,519.04	$1,994.23	$10,193.81	0.478	7,833.83	L
17	16	49	57,313.21	2.32	$3,093.62	$2,733.58	$63,142.73	$197.92	$163.64	0.1	$555.03	$916.69	$62,226.04	0.478	47,820.01	M

Problem Type

Multiclass Classification Type

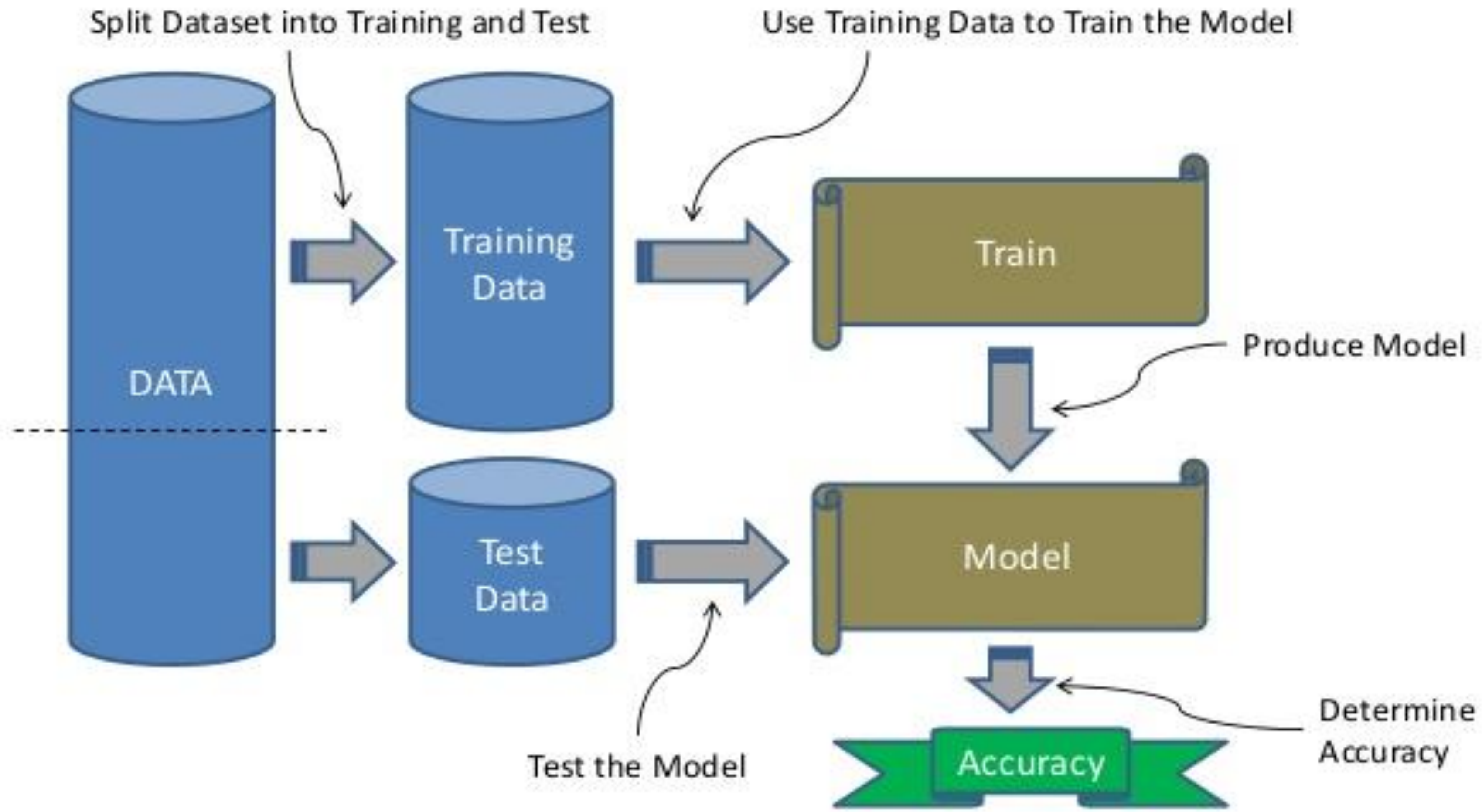
Low CLV- Range(100-30,000)
denoted by 0

Medium CLV- Range(30,000-75,000)
denoted by 1

High CLV – Range(75,000-*)
denoted by 2

It’s About Training

Machine Learning is about using data to train a model

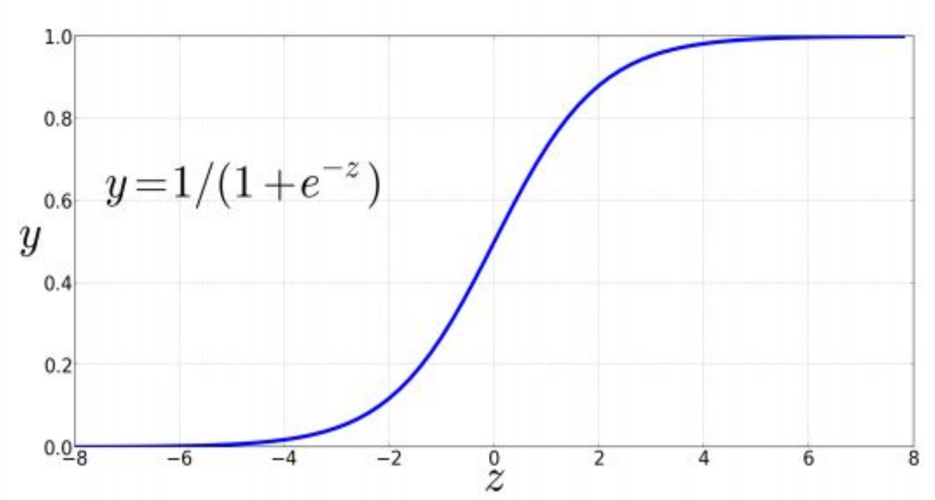


Machine Learning Models used

- Multiclass Classification using Random Forest
- Deep Learning – ANN
- Multiclass Classification using logistic regression

Multiclass Logistic Regression

- Logistic regression is a classification algorithm.
- **Logistic regression is designed for two-class problems**, modeling the target using a binomial probability distribution function. The class labels are mapped to 1 for the positive class or outcome and 0 for the negative class or outcome. The fit model predicts the probability that an example belongs to class 1.
- By default, logistic regression can be used for classification tasks that have more than two class labels, so-called multi-class classification
- **Hence for problems, that have more than one groups to classify are called <u>Multiclass Logistic Regression</u>.**



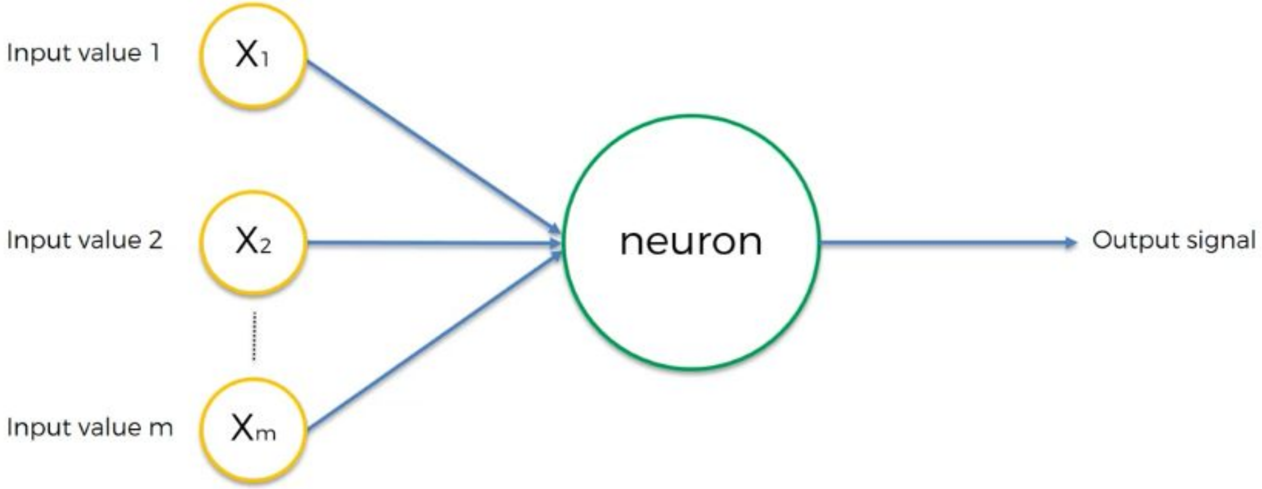
The mathematical function that essentially gives a binary output or 0 or 1. To do this, we can use the **logistic** or **sigmoid** function, which has the form:

ANN Using Deep Learning

- Deep learning is an artificial intelligence (AI) function that imitates the workings of the human brain in processing data and creating patterns for use in decision making.

- Deep learning is a subset of machine learning in artificial intelligence that has networks capable of learning unsupervised from data that is unstructured or unlabeled. Also known as deep neural learning or deep neural network.

- As a subset of machine learning, deep learning uses hierarchical neural networks to analyze data. Neuron codes are linked together within these hierarchical neural networks, similar to the human brain. Unlike other traditional linear programs in machines, the hierarchical structure of deep learning allows it to take a nonlinear approach, processing data across a series of layers which each will integrate subsequent tiers of additional information



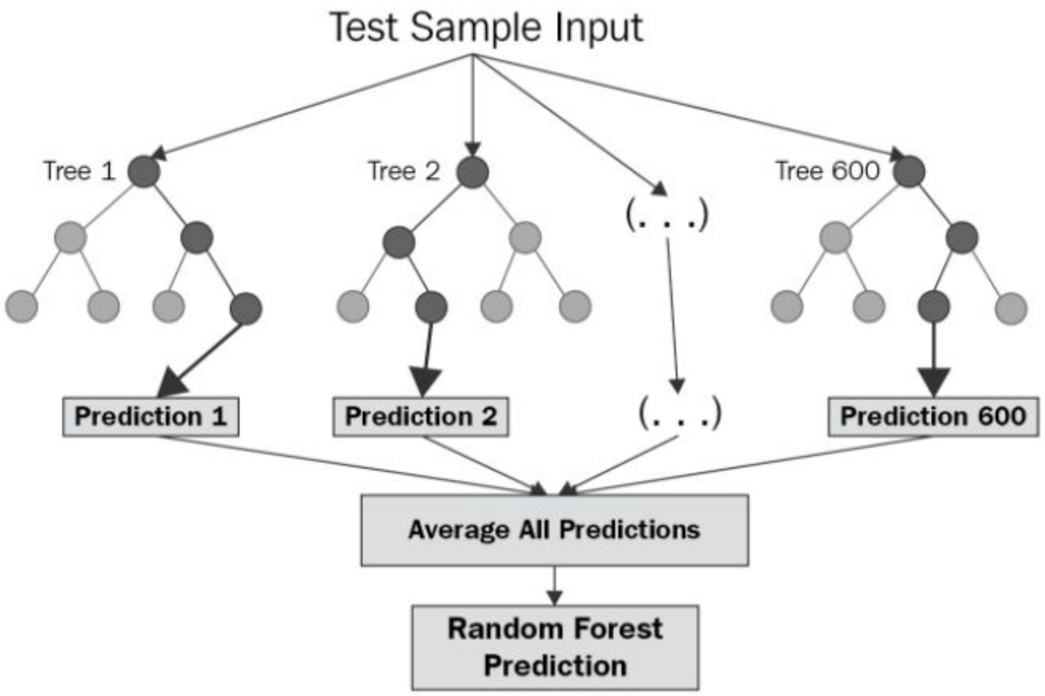
- The inputs here are independent variables, and all these are present for one single observation which we are training our model on. Also, these variables need to be either standardized (making sure they have a mean of zero and a variance one) or normalized making them to fit in about a range of values.
- This is done as all these values are added up or multiplied in a neural network, hence it will be easier to process then if they’re all about the same.
- Here, the output values can be continuous, it can be binary or it can be categorical variable (several output values)

Multiclass Classification using Random Forest

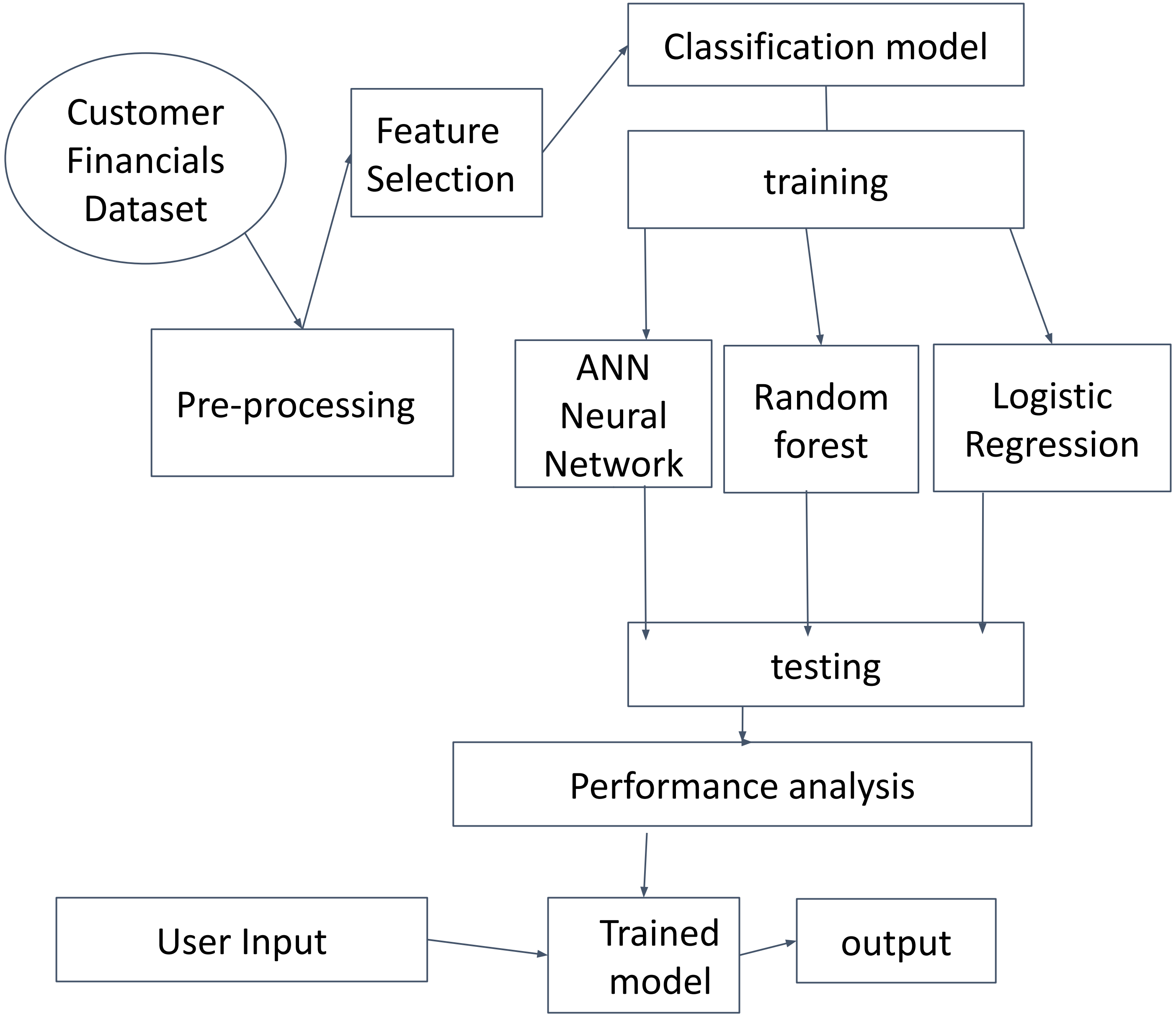
- **Random Forest Regression** is a supervised learning algorithm that uses **ensemble learning**
- Method for regression. Ensemble learning method is a technique that combines predictions from
- Multiple machine learning algorithms to make a more accurate prediction than a single model.
- A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.

To get a better understanding of the Random Forest algorithm, let’s walk through the steps:

1. Pick at random *k* data points from the training set.
2. Build a decision tree associated to these *k* data points.
3. Choose the number *N* of trees you want to build and repeat steps 1 and 2.
4. For a new data point, make each one of your *N*-tree trees predict the value of *y* for the data point in question and assign the new data point to the average across all of the predicted *y* values.



High Level Design



Customer Financials Dataset

Pre-processing

Feature Selection

Classification model

training

ANN Neural Network

Random forest

Logistic Regression

testing

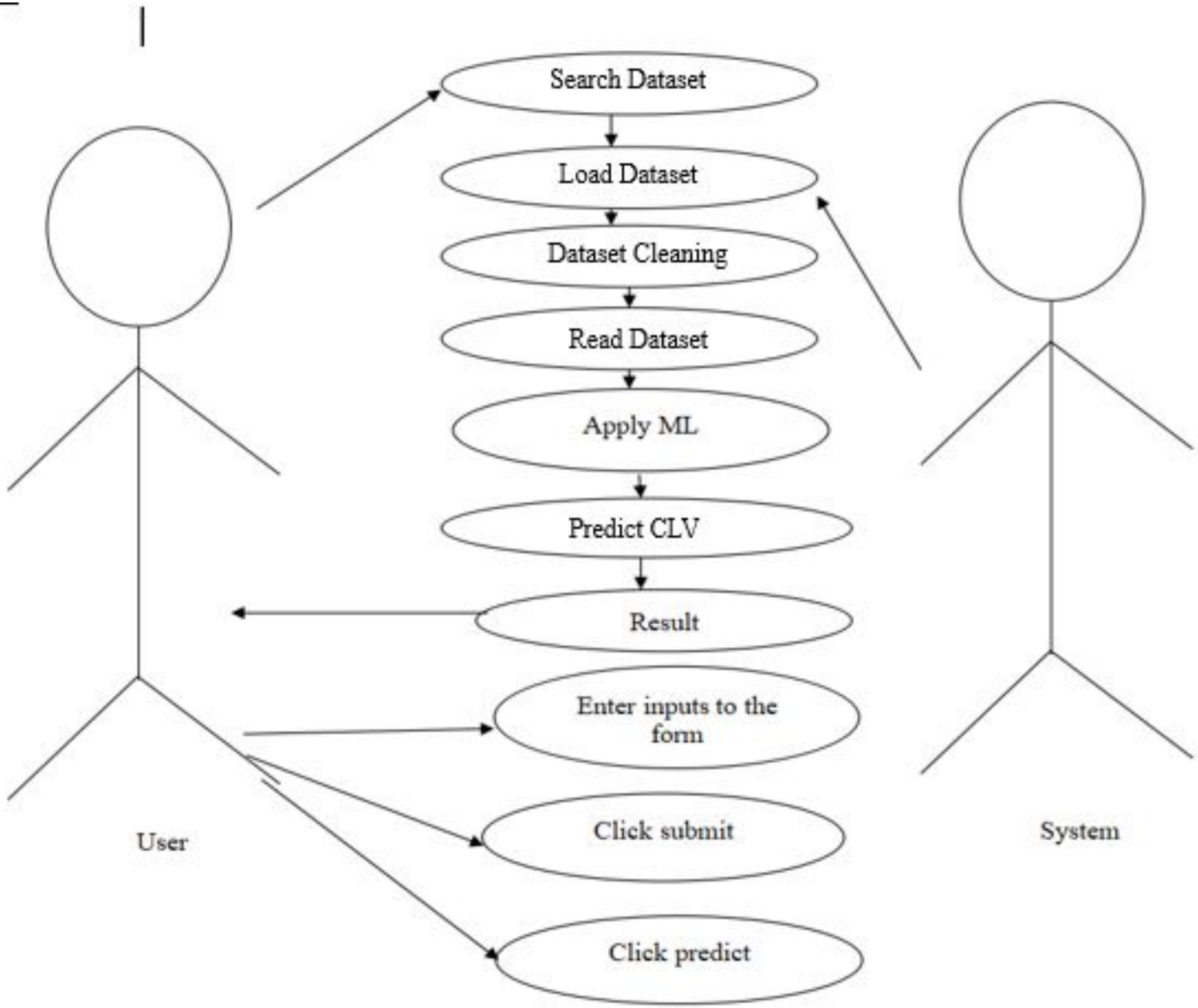
Performance analysis

User Input

Trained model

output

Use-Case Diagram



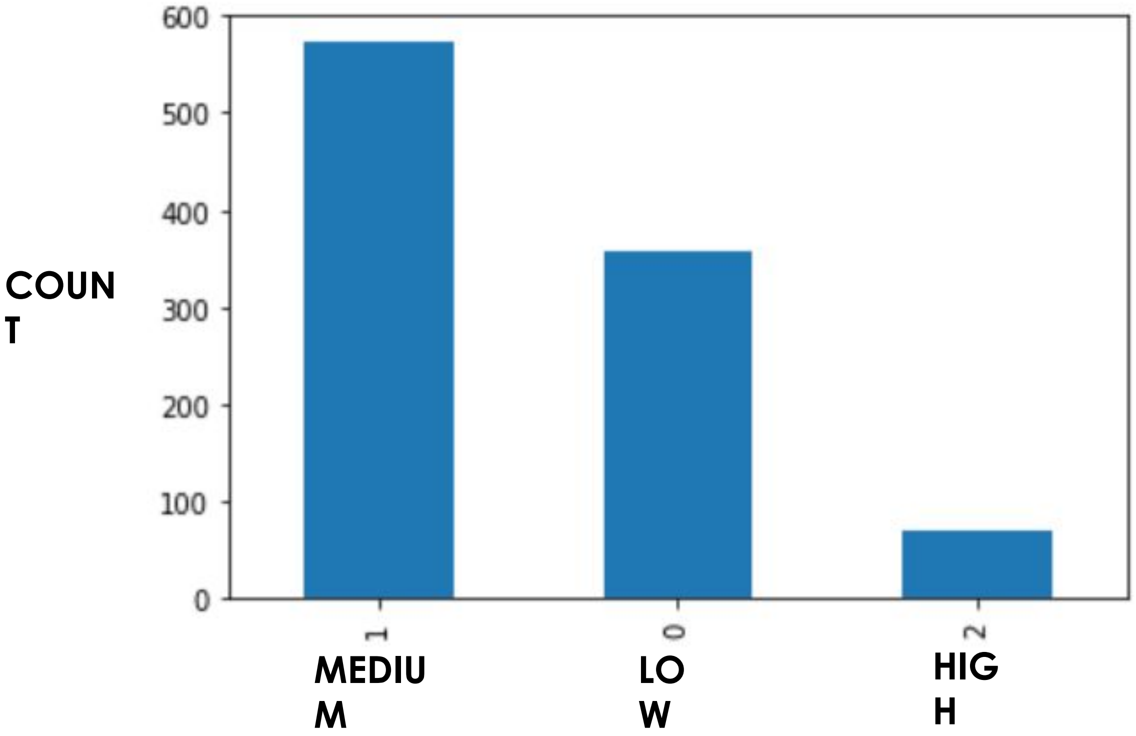
Results-(Model Accuracy)

- Logistic Regression -(87%)
- Random Forest -(99.3%)
- ANN Modelling -(97.3%)

Graphical Results

Count of Various CLV's

```
Out[55]:  <matplotlib.axes._subplots.AxesSubplot at 0xeb1841ce4
```

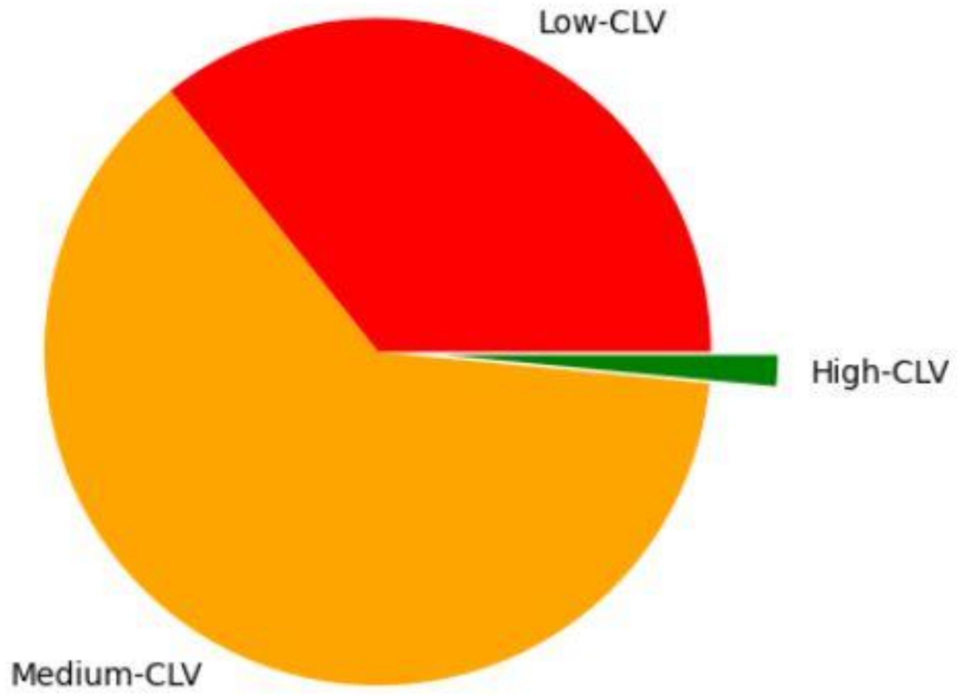


Various CLV values

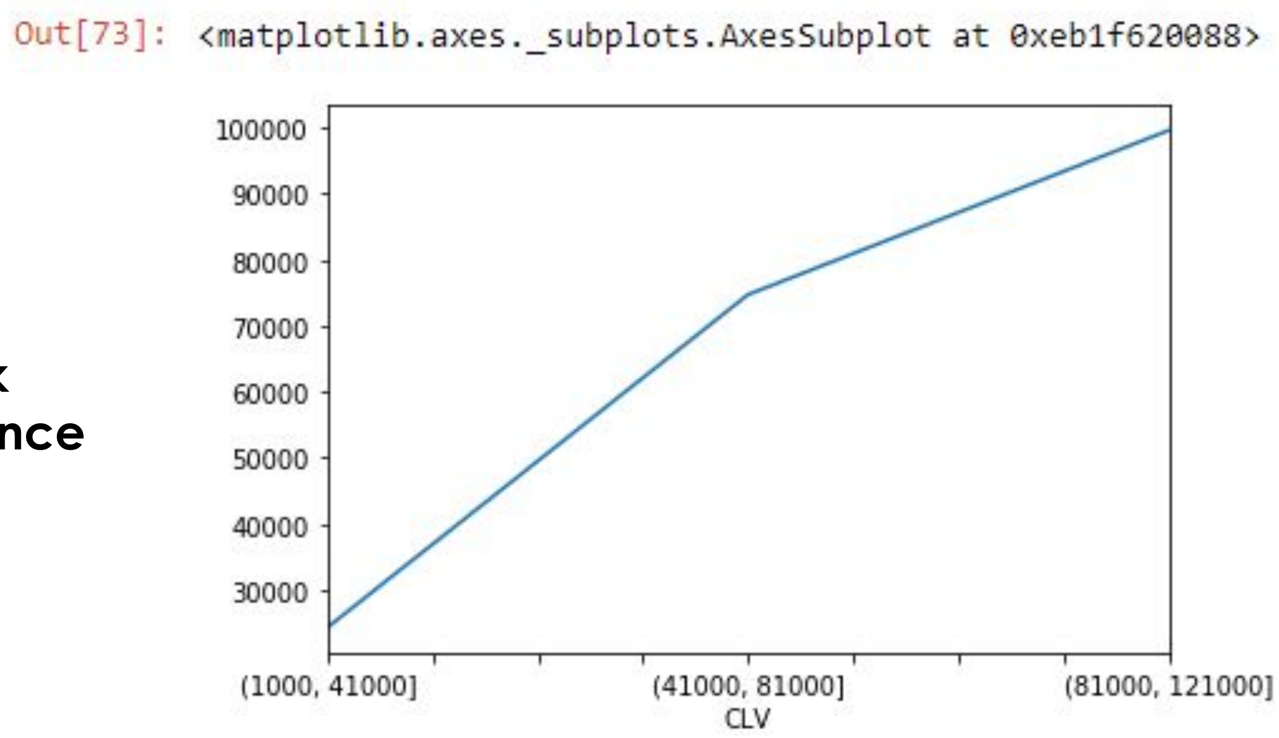
- Low CLV's between (300,400)
- Medium CLV's between (500-600)
- High CLV's between (10-100)

Existence of Various CLV's in a Sample

Various CLV Types Presenet in a Sample of 1000 Customers



Plot Between Bank Balance and CLV

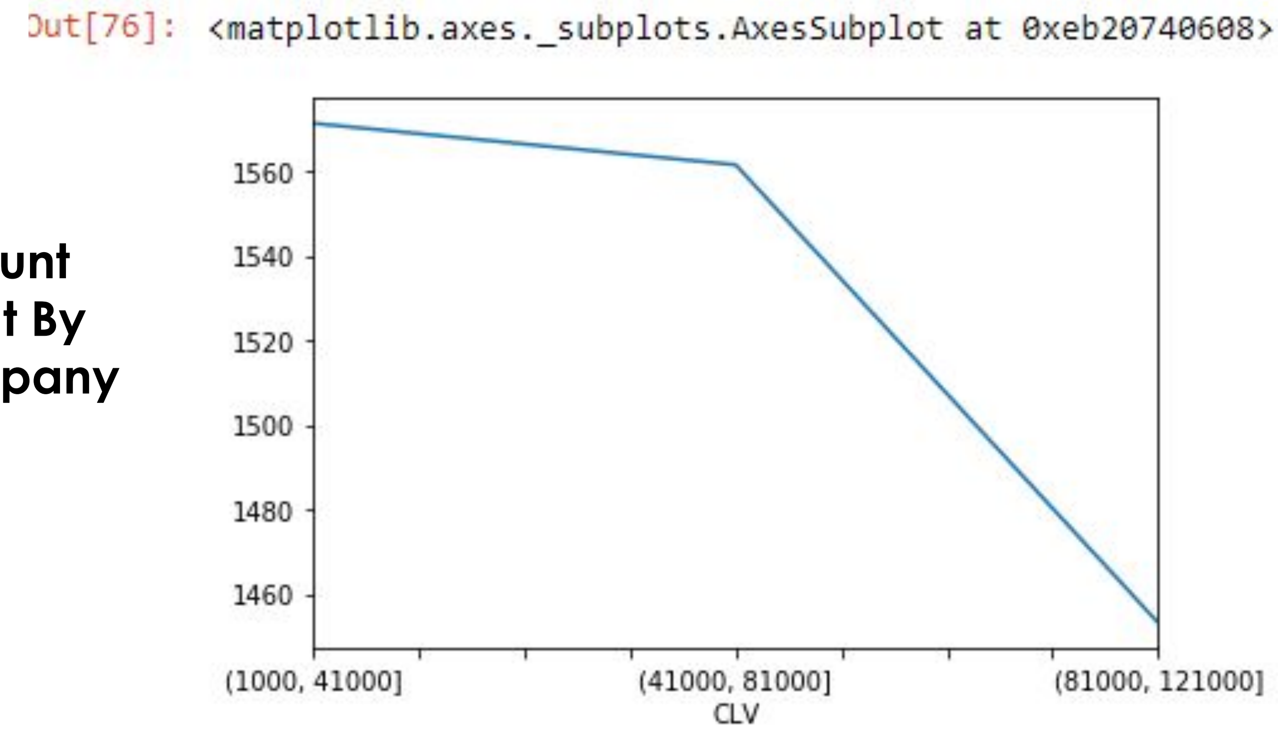


```
Out[73]: <matplotlib.axes._subplots.AxesSubplot at 0xeb1f620088>
```

Bank Balance

- Customers with **Higher Bank Balance** tend to have a **Higher CLV** value.

Plot Between CLV and Spent-Amount

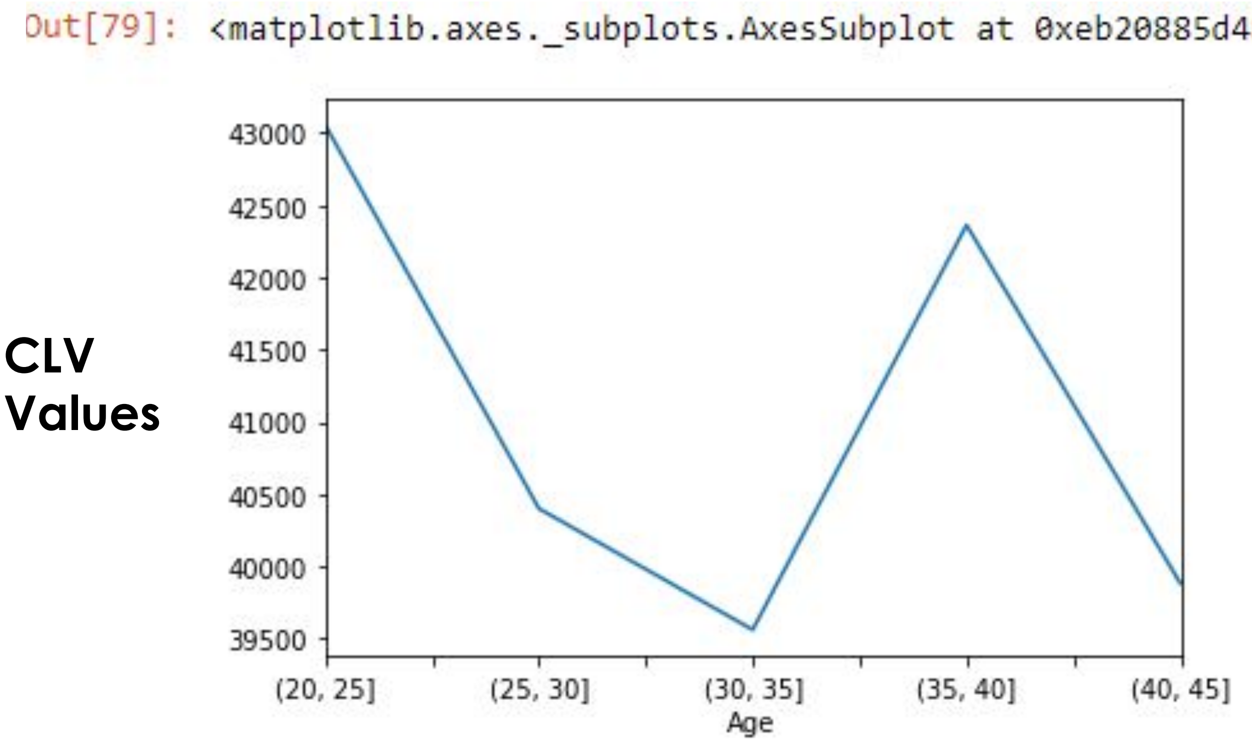


```
Out[76]:  <matplotlib.axes._subplots.AxesSubplot at 0xeb20740608>
```

Amount Spent By Company

High Valued Customers give more and take less

Plot Between Age and CLV

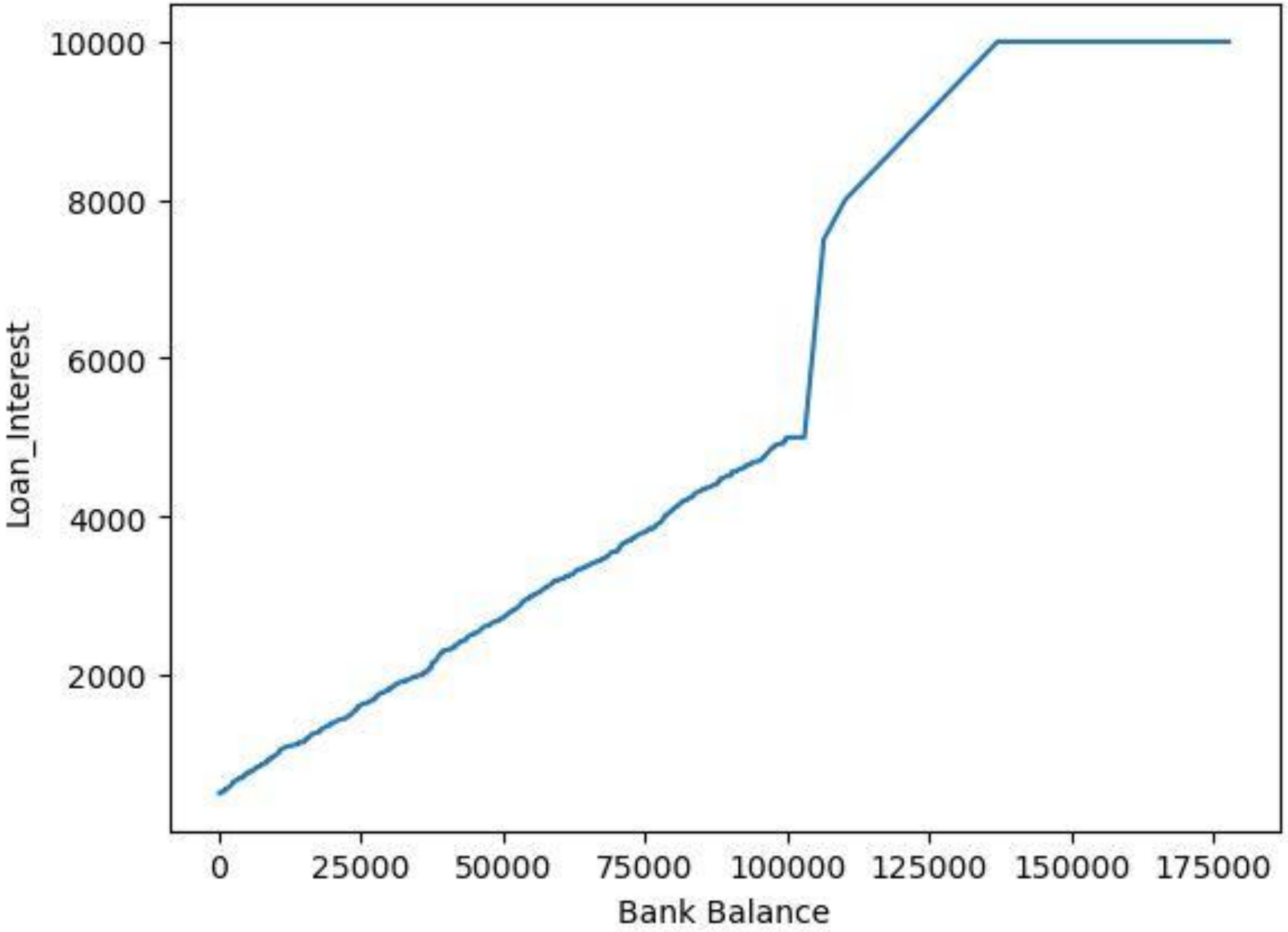


```
Out[79]: <matplotlib.axes._subplots.AxesSubplot at 0xeb20885d48>
```

CLV Values

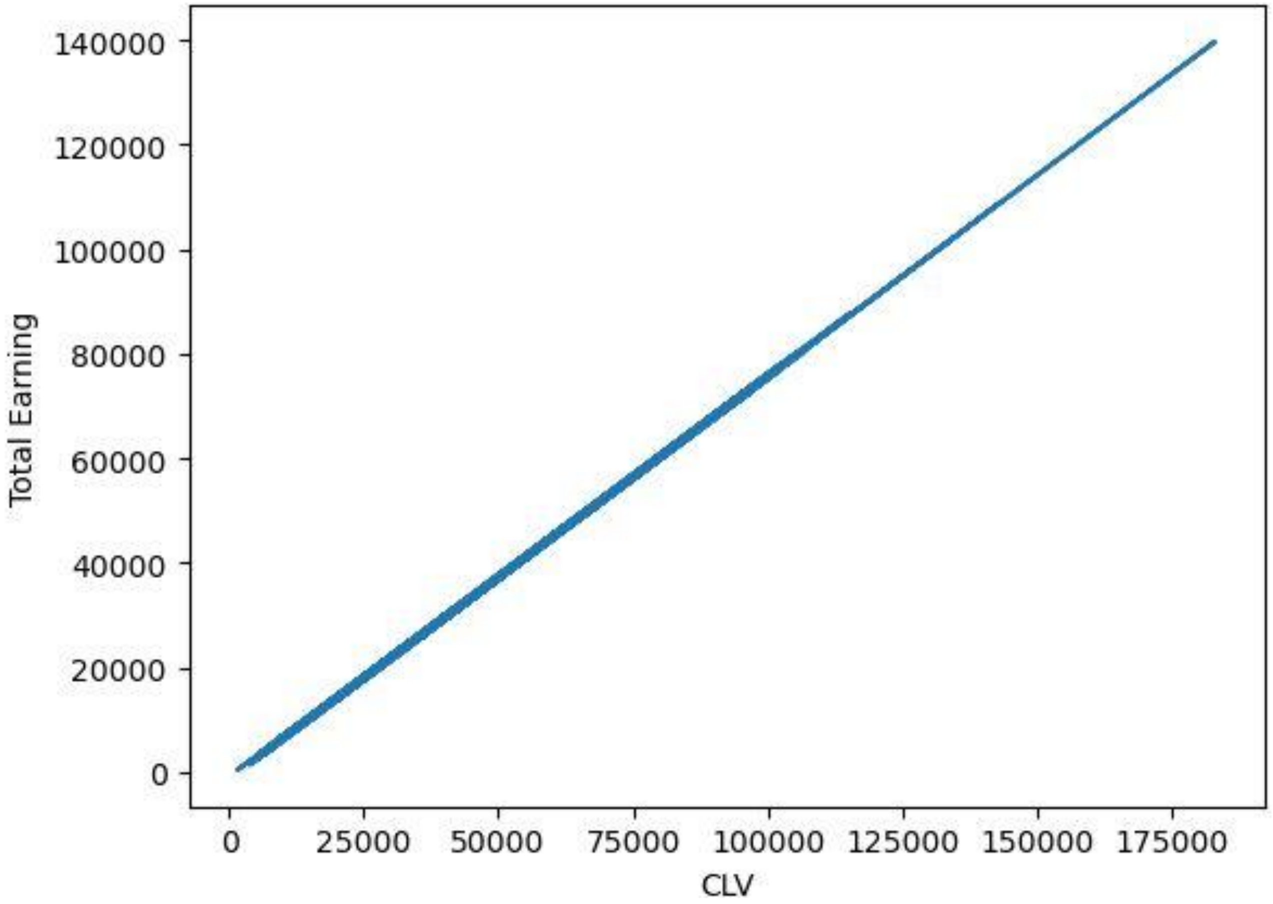
- People Within the age group of (30,35) have a much lower mean salary then those Younger and older

Plot Between Bank Balance and Loan Interest

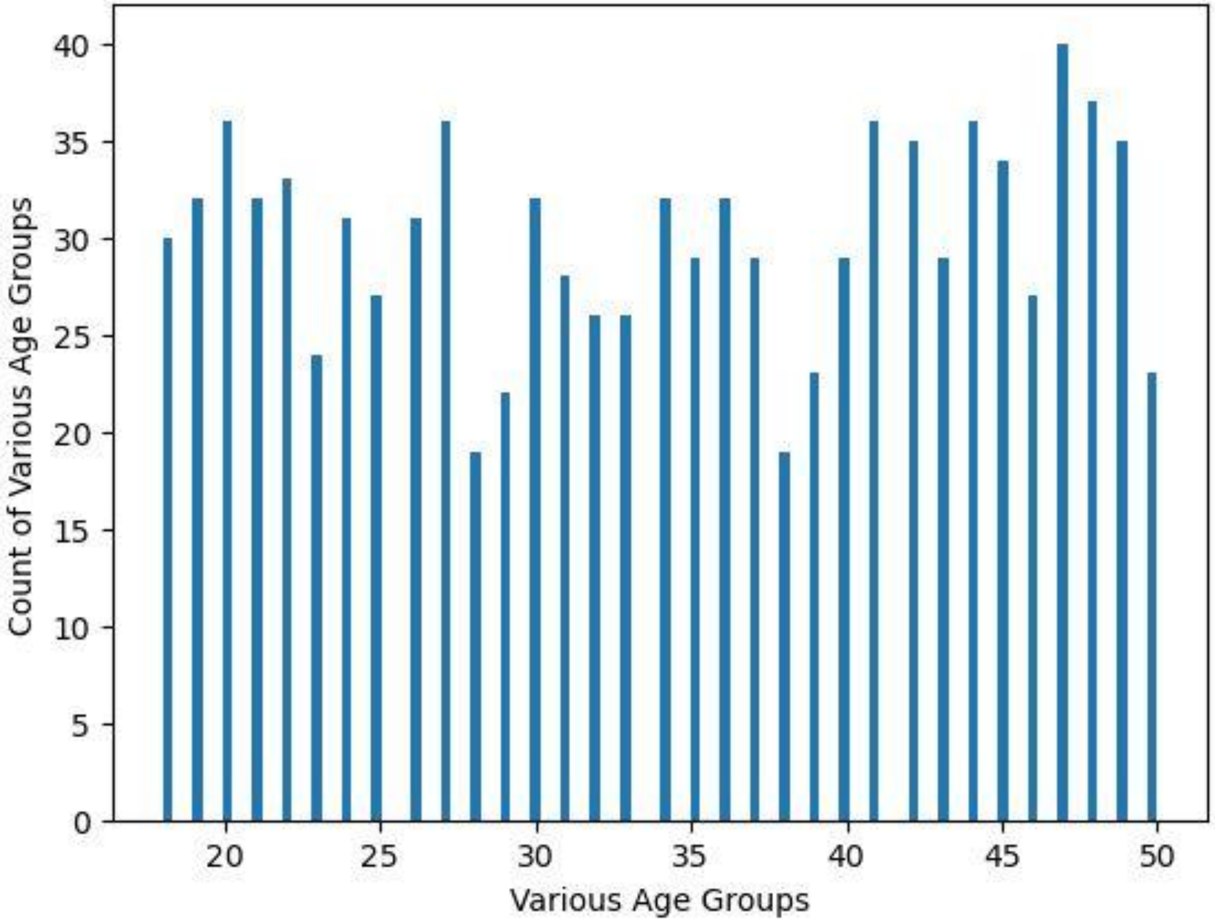


People with Higher Bank Balance tend to pay more Interest on Loans

Plot Between Total Earning and CLV

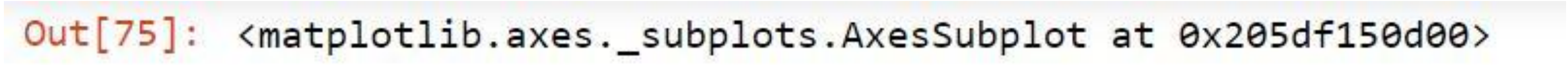


Histogram Plot of Count of Various Age Groups

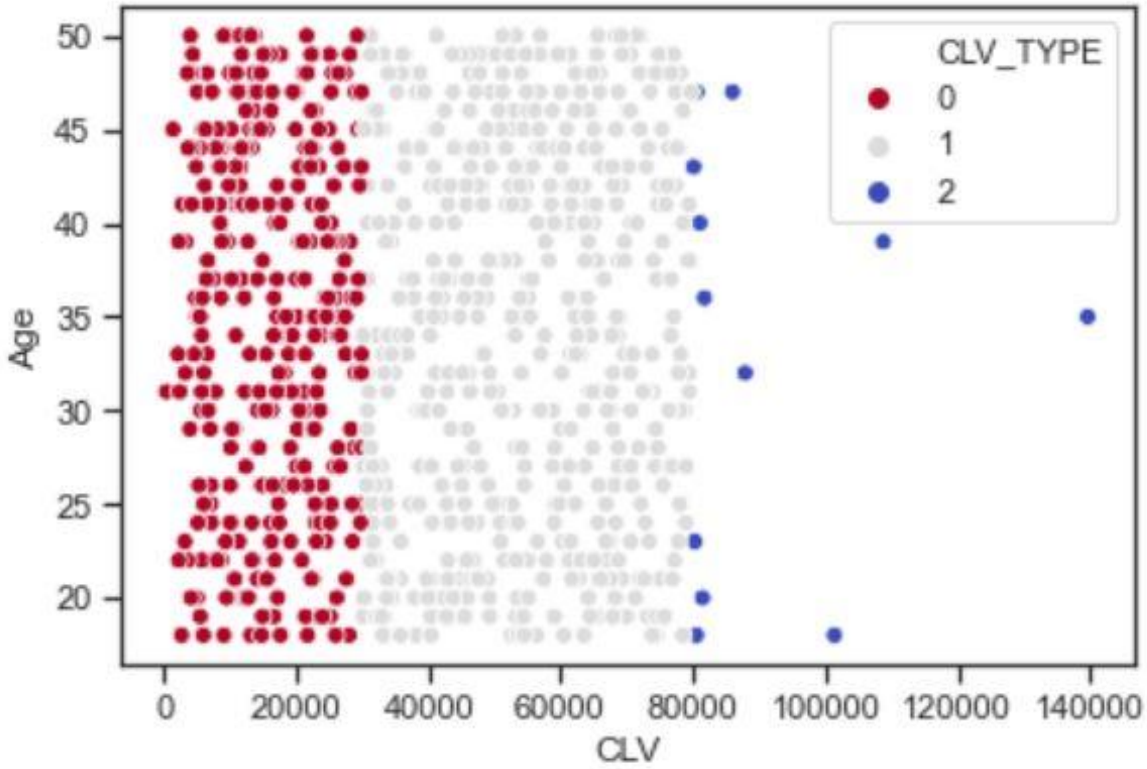


- Youngest Customer: 18 year old
- Oldest Customer: 50 year old
- Age Groups with the highest presence: b/w 45-50
- Age Groups with the lowest presence: b/w 18-20

Scatter Plot of Various Age groups and their CLV's

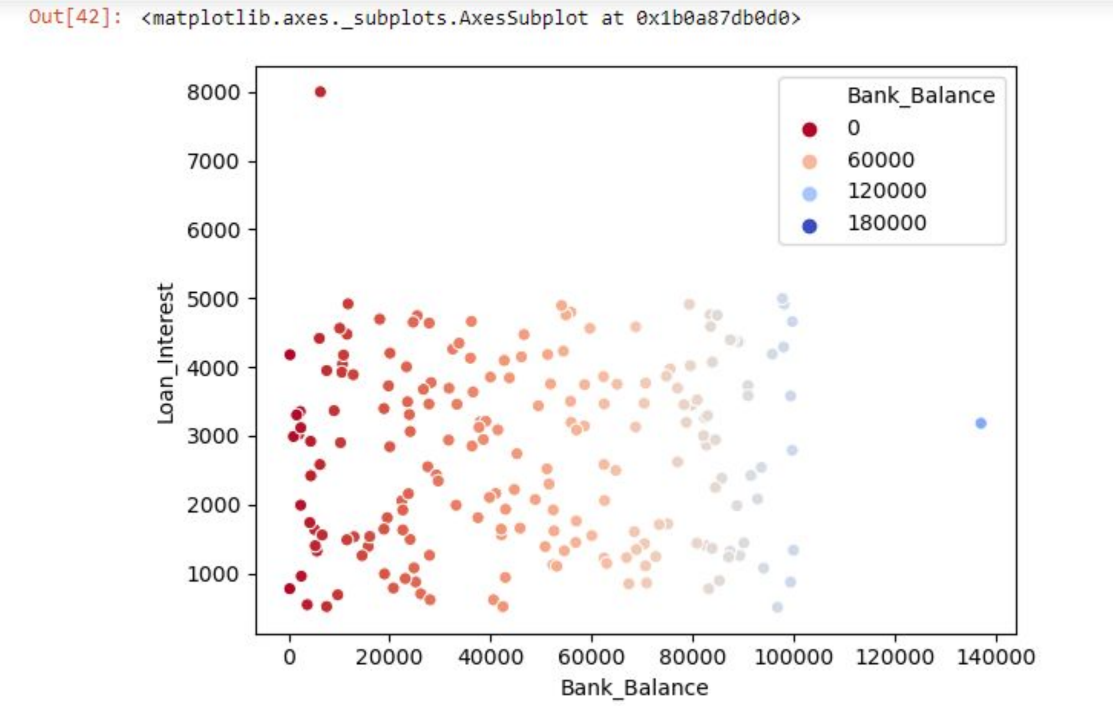


```
Out[75]:  <matplotlib.axes._subplots.AxesSubplot at 0x205df150d00>
```

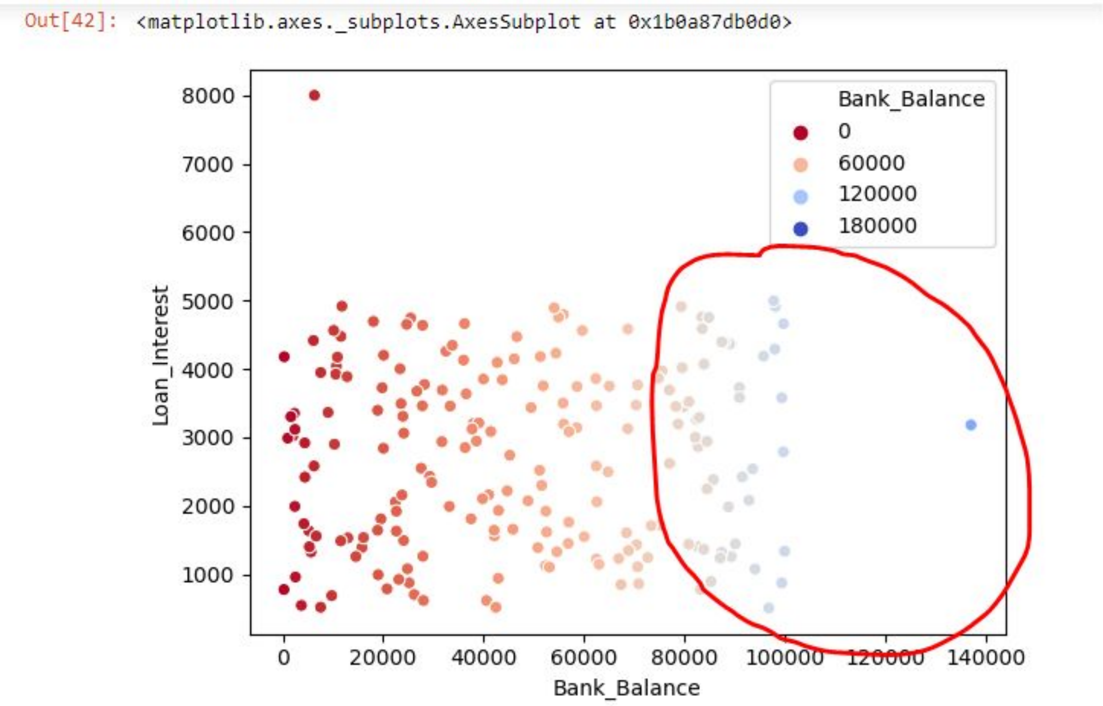


We have analyzed the various age groups and their respective CLV, and have agreed that the Value of CLV is evenly distributed across various age groups

Scatter Plot of Various Bank Balance and their Loan Interest Payment – The Points inside the Circle are the Customers to Target



```
Out[42]: <matplotlib.axes._subplots.AxesSubplot at 0x1b0a87db0d0>
```



```
Out[42]: <matplotlib.axes._subplots.AxesSubplot at 0x1b0a87db0d0>
```

B.M.S. COLLEGE OF ENGINEERING
ESTD. 1946
BANGALORE - 560019

Thank You