Introduction to Marketing Analytics

(MANM317)

Amol Dixit (6562752)

Master of Science in Business Analytics

University of Surrey

Words: 2580

Faculty of Arts and Social Sciences

University of Surrey

Guildford, Surrey, GU2 7XH, UK

March 2019

Executive Summary

This report summarises the findings of the Customer Management project within the Marketing Analytics Department of our bank which holds a prominent position in the spectrum of retail banking. The major objective of this project is to find out which marketing activity retains the customers with the bank, and generates high profitability. For this a wide set of data was taken and analysed which included,

(i) Personal details about customers; such as gender, profession, income, education etc.

(ii) Customer-Bank relationship details; such as percentage of loans & savings with our bank, branch number, relationship duration etc.

(iii) Results of a specially drafted survey with questions relating to customer's satisfaction, loyalty and value.

(iv) Customer Lifetime Value for a set of customers.

This data helped in doing a descriptive analysis and find out which customer attributes are linked to high profitability over a lifetime. Apart from this a predictive analysis is also done to find out the overall profitability a customer would bring in based on the other factors, with certain level of confidence. This helps in zeroing on the features which drive CLV the most and the model which helps in predicting the CLV assists in planning our interactions with the customers.

The following report describes the various data analysis models used and the reasons why they were chosen and also compares the results obtained from them and finally makes recommendations with the best suited model.

The models are built using a bottom-up approach, starting from the single variable, and then adding complexities to it, so as to enhance the predictive accuracy while maintaining the significance of variables.

Also, the models are designed on approximately 70% of the data-points and validated on the remaining 30%, this helps us in proving the robustness of the model, and its applicability to new-unseen data. These scales are chosen as per conventional prevalence.

Table of Contents

Executive Summary 2
Table of Contents 3
Introduction 4
Premises / Theoretical Foundations 5
Methodology 7
Data Analysis 14
Model A 14
Model B 16
Model C 16
Conclusion 18
Recommendations 19
References 20
Appendices 21
Appendix A: Scatter Plots 21
Appendix B: Correlation Matrix 25
Appendix C: Model C and higher powers of income and loan 26
Appendix D: Loyalty based models 26
Appendix E: Model B 28
Appendix F: IBM-SPSS Modeller Output File 29

Introduction

Customer Lifetime Value (CLV) is a very important factor for any company when deciding its long-term goals. It is the net profit from the entire relationship with a customer that a company can expect to have. It is purely a marketing concept, but is extensively used in defining the marketing strategies and budget allocations not just on quarterly basis but on longer durations ranging from a few months to several years. The CLV prediction model can be in the form of a simple formula defining net value or a complex data model which bases its results on numerous factors of the customer and the customer-client relationship.

The data set has records for 737 customers, from 6 different bank branches, with details about the person's employment, gender, age, level of education, income category, the percentage of savings and loans with our bank, when they opened their account **and their CLV**. Also, the customers have answered questions to a survey on value, satisfaction and loyalty; the missing values from this survey are imputed. The training and test data set is chosen by random sampling using Bernoulli distribution (Sean X. Chen and Jun S. Liu, 1997) and is widely used for probability distribution of random variables.

Features from all three groups, viz, personal, customer-bank relationship, and survey, were used to build models, they were also modified and grouped with each other to find better predictions. Correlations, mediators, control variables, higher powered functions, logarithmic variations, exponential relationships, Weighted Least Squares (WLS) methods and variable interactions were employed consideration and models were designed using combinations of these.

Premises / Theoretical Foundations

This section is divided into several labelled parts to discuss about the different areas which the project touches.

(A) The major objectives of marketing is to create, capture and maintain customer value, by:

- **Avoiding becoming irrelevant** on the market. A company´s profitability is based on creating offerings that serve unmet consumer needs (eg: UBER v/s conventional taxis).

- **Maximising long-term profits**, companies with a high customer focus achieve better long-term profits. (Best Roger J., 2012). eg: the lifetime revenue of a corporate purchaser of commercial aircraft can be billions of dollars.

Figure.1: *Examines the top and bottom performers in the American Customer Satisfaction Index database (+200 companies) and presents a stock price index for these companies over 10 years based on the customer satisfaction level*

(B) The retail banking CLTV is based highly on the income and the loyalty of customers for a bank (Limam M, 2012). Also, the cost of acquiring such a customer is much lower than that of latter (Deely M., 2011), so ensuring that a customer remains loyal provides double benefits by reducing costs and increasing value.

(C) The usage of regression for analysing CLV is based on the work described by Prof. Jesper Ryd´en, who also suggested that Cook's distance for outliers and ANOVA tables are important to look at (Ryd'en,J., 2014).

(D) When R^2 is similar for both, test and validation datasets, the model can be considered as unbiased. (R. Arboretti Giancristofaro, L. Salmaso, 2003).

(E) WLS allows us to obtain a weighted least-squares model (IBM, 2018). By this the data points are weighted by the reciprocal of their variances, due to which observations with large variances have less impact on the analysis than those with small variances. This statistical technique is used in this project to build and improve the models.

With a balanced theoretical base, spanning the domains of marketing, statistics and analysis, we move on to the methodology employed in this project.

Methodology

The project was done by following the CRISP-DM (Cross-Industry Standard Process for Data Mining) process model, and moving forward/backward was done several times upon finding a newer insight in the data or to try another approach. New variables whenever introduced were taken along the same processes and the analysis was modified.

Figure.2: *The CRISP model for Data Mining (Jensen, K, Source: www.wikimedia.com)*

The **first** thing done as part of this project was to go through the data and features, detect the missing values, outliers etc. Some essential points were noted in this step which were kept in consideration while building the models, like,

i. 90% Customers agreed more to Sat3_1 than they did to the other two.

ii. There were some misnomers in gender (=0), and missing values in gender, professions, education etc. Models were tried by dealing with these missing values.

iii. Outliers in the loan variable (eg 1%): models were tested by both including and excluding.

iv. The questionnaire responses had no missing values, but did have out of scale values. By discussions with the data provider it was found that values were imputed.

Next, a correlation analysis was done between the variables and the target field, CLV, (see **Figure.4**) and it was seen that CLV is not highly linearly dependent on any variable, but most depended on *income*. Also, some survey results with high correlations could be combined as one for simplicity and avoid multi-collinearity (see Fig 4).

Although, the correlation between loyalty, value & satisfaction and the target field (CLV) is very small, all of them will be considered for the analysis, as one of the key aspects of customer value is the **perception of benefits**. The importance of the perceived benefits is specific to each customer and needs to be quantified to serve as an input for managerial decisions.

		SMEAN(val1)	SMEAN(val2)	SMEAN(val3)	SMEAN(sat1)	SMEAN(sat2)	SMEAN(sat3)
Spearman's rho	resp_no	0.035	0.037	0.051	0.039	0.029	-0.009
	SMEAN(val1)	1	0.788	0.701	0.465	0.395	0.499
	SMEAN(val2)	0.788	1	0.758	0.489	0.444	0.556
	SMEAN(val3)	0.701	0.758	1	0.57	0.488	0.62
	SMEAN(sat1)	0.465	0.489	0.57	1	0.807	0.565
	SMEAN(sat2)	0.395	0.444	0.488	0.807	1	0.616
	SMEAN(sat3)	0.499	0.556	0.62	0.565	0.616	1

Figure.3: *Sample Correlation matrix (complete chart in Appendix B)*

Correlation with CLV			
profess	**-0.101**	SMEAN(val3)	**0.021**
gender	**-0.015**	SMEAN(sat1)	**0.016**
educatio	**0.029**	SMEAN(sat2)	**-0.018**
SMEAN(loy1)	**0.107**	SMEAN(sat3)	**0.041**
SMEAN(loy2)	**0.144**	SMEAN(sav)	**0.044**
SMEAN(loy3)	**0.101**	SMEAN(loans)	**0.091**
SMEAN(loy4)	**0.125**	SMEAN(long)	**0.011**
SMEAN(loy5)	**0.005**	SMEAN(age)	**0.105**
SMEAN(loy6)	**0.068**	SMEAN(income)	**0.195**

Figure.4: *Correlation between CLV (target) and other variables.*

We get several insights from Fig.4, such as *sat3* having a greater effect on CLV than the other satisfaction-questions. And, we can say with some certainty that *sat_2* is a negative question, as its correlation with CLV is <0. Also, when looking at loyalty, it is evident that *loy_5* does not affect CLV and can be discarded, and since other loyalty related questions have similar values they can be averaged and considered as one.

The base model with *income* is summarised below.

$$y = f(SMEAN(income))$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.208[a]	.043	.042	4397.73237

a. Predictors: (Constant), SMEAN(income)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	4282.021	453.534		9.441	.000
	SMEAN(income)	813.340	141.573	.208	5.745	.000

a. Dependent Variable: CLV

Figure.5: *Model Summary and Coefficient table from the base model*

After this analysis based on correlations, scatter plots were made to visually analyse the dependence of CLV on various variables, and to choose subsequent independent variables. For example below scatter plot shows the relation between the *income* and *CLV*, we can see that there exists a slight linear relation between them. For highest income groups though it drops, probably due to customers looking for more luxurious services.

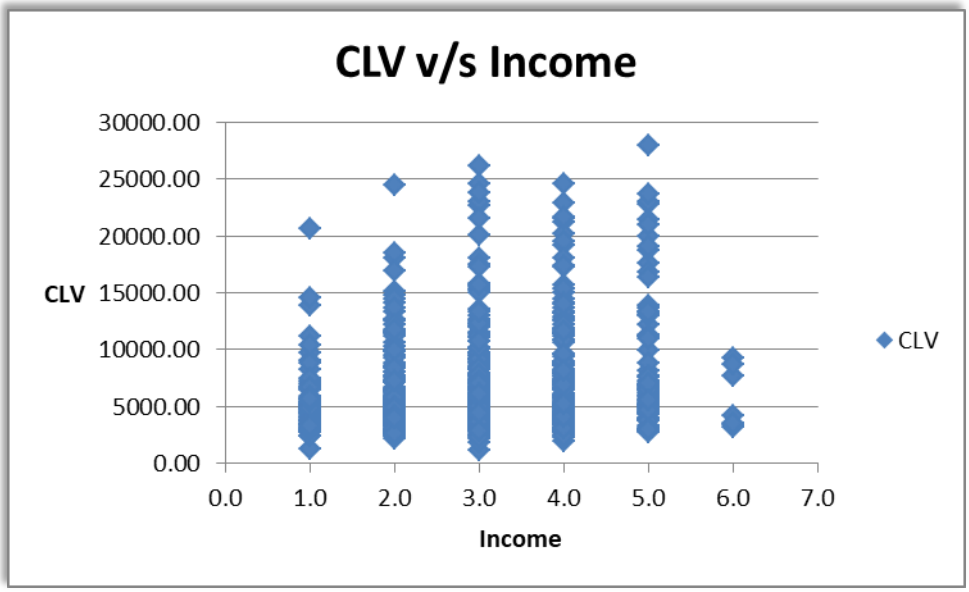

Figure.6: *Scatter plot between Income and CLV*

Another variable on which CLV shows a diverse and varying dependence on is loans. When we make a scatter plot between CLV and loans and plot a trendline, we get a higher order term with an R^2 of ~14%.

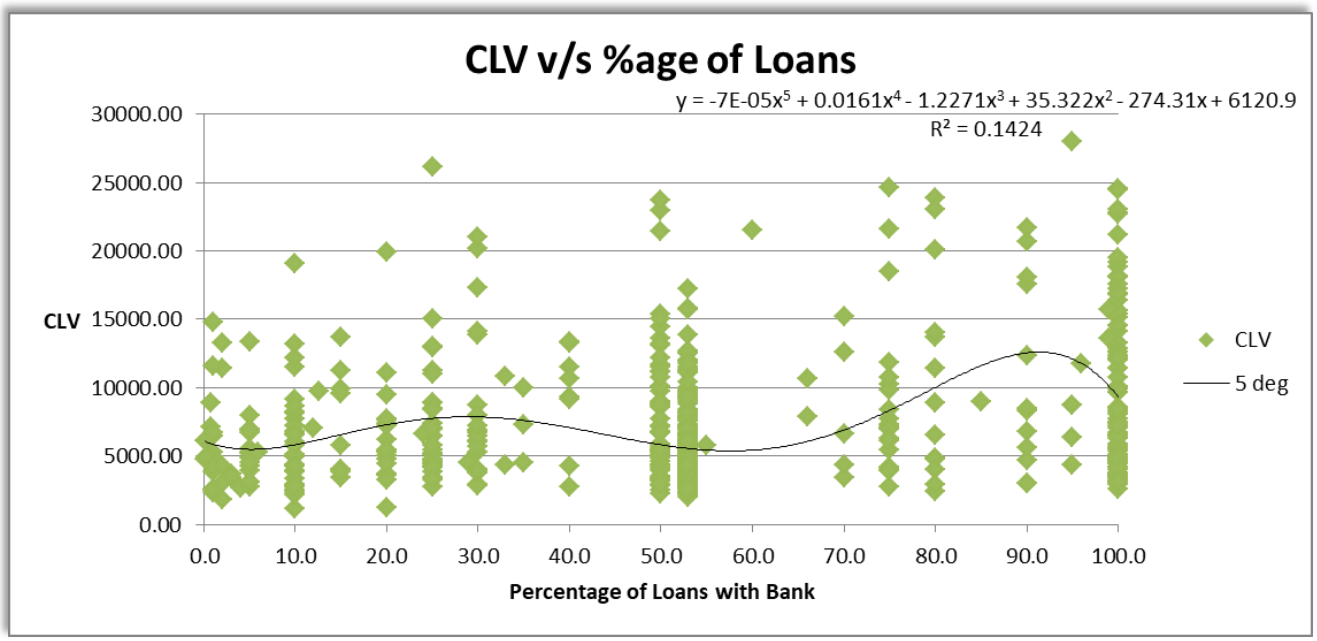

Figure.7: *Scatter plot between Loans and CLV with trendlines*

This analysis helps us in building the model up from the single income variable. Scatter plots for CLV v/s other relevant variables are provided in the Appendix A below.

The models are judged on the basis of how well they explain (R^2) the CLV, and to validate this, the dataset is partitioned into Training (70%) and Validation (30%) sets by using Bernoulli distribution, as below:

```
COMPUTE Training =RV.BERNOULLI(0.7).
EXECUTE.
```

Figure.8: *Partitioning the dataset.*

The next step came from the results of visual analysis, and newer variables were generated, starting with derivatives of *loyalty*, and the models are compared to choose the best set of features. R^2 is significantly higher by excluding *loy_5* which establishes our previous statement. Also, the p-values (see Fig 9) show that the coefficients are significant at every conventional level of significance. Other *loyalty* based models are shown in Appendix D.

Model Summary[b,c]

Model	R	Training = 1 (Selected)	Training ~= 1 (Unselected)	R Square	Adjusted R Square	Std. Error of the Estimate
1		.150[a]	.201	.023	.021	4609.13420

a. Predictors: (Constant), loy_mea_12346

b. Unless noted otherwise, statistics are based only on cases for which Training = 1.

c. Dependent Variable: CLV

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3173.327	1055.841		3.005	.003
	loy_mea_12346	732.809	209.981	.150	3.490	.001

a. Dependent Variable: CLV

b. Selecting only cases for which Training = 1

Figure.9: *Model by using all the loyalty pointers except 5*

Several models were subsequently made with different combinations of variables, and by generating variables through different logics.

For example,

i. means of survey question sets
ii. higher power terms of income and loans
iii. interaction variables between features
iv. exponential functions
v. logarithmic functions

Loy_Mean_1_6	loans_power2	ques_comb	sat1_val3_sum
Loy_Mean_1_3	loans_power_3	que_pro	sat1_val3_avg
Val_Mean_1_3	loans_power_4	que_edu	sat_1_2_Val_all_Div_2
Val_Mean_1_3_with2	loans_power_5	que_gen	sat_1_2_Val_all_Div_5
Sat_Mean_1_3	age_income	que_bra	loy_mea_12346
income_power2	sat_val	que_age	loy_1_3_exp
income_power3	prof_edu	que_inc	pro_long
income_power4	sav_long	val_inv	val_sat_avg

Figure.10: *The numerous variables which were created and tested*

After tests with several combinations we arrive at a model with *exp(loyalty)* and higher power terms of *income* & *loans* (see Fig. 11). Age is also seen as a major factor here. The value of R^2 (14.5%) is higher than the one obtained from the previous model (~2%) and all variables are significant at a 5% significance level except $loans^2$ and $income^3$ which are significant at 10% and 15% significance level respectively.

Model Summary[b,c]

Model	R Training = 1 (Selected)	R Training ~= 1 (Unselected)	R Square	Adjusted R Square	Std. Error of the Estimate
1	.393[a]	.475	.154	.145	4307.78274

a. Predictors: (Constant), SMEAN(age), loans_power2, loy_1_3_exp, income_power3, income_power2, loans_power_3

b. Unless noted otherwise, statistics are based only on cases for which Training = 1.

c. Dependent Variable: CLV

Coefficients[a,b]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-2127.446	2257.830		-.942	.346
	loy_1_3_exp	3411.207	1436.108	.096	2.375	.018
	income_power2	348.583	142.137	.565	2.452	.015
	income_power3	-40.967	25.772	-.365	-1.590	.113
	loans_power2	-.655	.373	-.416	-1.758	.079
	loans_power_3	.010	.004	.698	2.947	.003
	SMEAN(age)	34.846	16.011	.089	2.176	.030

a. Dependent Variable: CLV

b. Selecting only cases for which Training = 1

Figure.11: *Sample Model using the exponent function*

Also, another thing noticed during the directed trial and errors being performed was that **interaction variables**, gave high accuracy results but the coefficients were not significant. And the variables such as *loan*, *saving* and *long* had a large variance in their scales owing to the very nature of these variables, and this was causing the model to not be fitted properly. So to combine the above two situations for our best benefit, a **Weighted Least Squares** approach was used to include the interaction variable as part of the WLS Weight. Upon checking, it was seen that the interaction between saving and duration of relation gave the best results so far. Its application on a data model is shown below.

Model Summary[b,c,d]

Model	R	Training = 1 (Selected)	Training ~= 1 (Unselected)	R Square	Adjusted R Square	Std. Error of the Estimate
1		.445[a]	.480	.198	.189	198226.1676

a. Predictors: (Constant), loy_mea_12346, loans_power2, SMEAN(age), SMEAN (income), loans_power_5, loans_power_3

b. Unless noted otherwise, statistics are based only on cases for which Training = 1.

c. Dependent Variable: CLV

d. Weighted Least Squares Regression - Weighted by sav_long

Figure.12: *Sample Model using the WLS approach*

Which explain ~20% of variation in CLV as compared to the below model with the same variables but not following the WLS approach explaining ~16%.

Model Summary[b,c]

Model	R	Training = 1 (Selected)	Training ~= 1 (Unselected)	R Square	Adjusted R Square	Std. Error of the Estimate
1		.402[a]	.483	.162	.152	4288.27913

a. Predictors: (Constant), loy_mea_12346, loans_power2, SMEAN(age), SMEAN (income), loans_power_5, loans_power_3

b. Unless noted otherwise, statistics are based only on cases for which Training = 1.

c. Dependent Variable: CLV

Figure.13: *Same Model without the WLS. Lower R^2value*

This model is further tuned and the best stats are then taken forward for recommendations. Another thing which was tried while doing the modelling was using **control variables** (such as gender, education, profession), to check the inter-dependence of features. For instance, it is very likely that *satisfaction* or *loyalty* is dependent on the *gender* of a customer, where one gender is loyal to the bank while the other typically isn't. Such control tests were done on models and seen if any such viable combinations affected the final model, as then then we could split the data accordingly and make separate models for each split.

Data Analysis

In this section of the report some of the models are discussed in detail and are evaluated based on various parameters.

Model A

The first model is based on variables in Fig. 14 and includes exponential and higher power terms. Although, it's a complex model it explains the data well as seen from R^2 in Fig 15.

Variables Entered/Removed[a,b]

Model	Variables Entered	Variables Removed	Method
1	sat1_val3_avg, SMEAN (income), SMEAN(long), loans_power_3, SMEAN (age), loy_1_3_exp, Sat_Mean_1_3, income_power4, SMEAN (loans), loans_power_5, income_power2, loans_power2[c]	.	Enter

a. Dependent Variable: CLV

b. Models are based only on cases for which Training = 1

c. Tolerance = .000 limit reached.

Figure.14: *Variables used in the first described model*

Another good thing about this model is that results obtained from the model are quite similar for both the Training and Test data, which shows that the **model is unbiased** and effectively represents the data, and that **approximately 18.3%** of CLV is explained by the model.

Model Summary[b,c]

Model	R Training = 1 (Selected)	R Training ~= 1 (Unselected)	R Square	Adjusted R Square	Std. Error of the Estimate
1	.427[a]	.498	.183	.164	4259.25079

a. Predictors: (Constant), sat1_val3_avg, SMEAN(income), SMEAN(long), loans_power_3, SMEAN(age), loy_1_3_exp, Sat_Mean_1_3, income_power4, SMEAN(loans), loans_power_5, income_power2, loans_power2

b. Unless noted otherwise, statistics are based only on cases for which Training = 1.

c. Dependent Variable: CLV

Figure.15: *The comparable values of R for both datasets*

The coefficients table (Fig.16) shows the variables and their standardised and unstandardized coefficients along with standard errors. Along with this it shows the critical value (Column t) obtained from the t-test and the p-value (Column Sig.). Typically a variable with p-value < 0.05 can be considered significant.

The coefficients can be interpreted as:

Keeping all other factors constant, if the SMEAN(age) increases by 1 unit (year) the CLV shall increase by 38.039 units (GBP) on average. And this estimation is highly significant.

Coefficients[a,b]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	60.762	2968.913		.020	.984
	Sat_Mean_1_3	-40.813	209.557	-.012	-.195	.846
	SMEAN(loans)	331.257	125.930	1.847	2.630	.009
	SMEAN(long)	-1.170	7.510	-.006	-.156	.876
	SMEAN(age)	38.039	15.911	.097	2.391	.017
	SMEAN(income)	-2795.790	1678.141	-.709	-1.666	.096
	income_power2	927.396	418.804	1.503	2.214	.027
	income_power4	-13.883	6.111	-.666	-2.272	.023
	loy_1_3_exp	3725.571	1557.558	.105	2.392	.017
	loans_power2	-13.879	4.327	-8.820	-3.207	.001
	loans_power_3	.166	.047	11.203	3.509	.000
	loans_power_5	-5.536E-6	.000	-3.913	-3.401	.001
	sat1_val3_avg	-137.101	222.394	-.040	-.616	.538

a. Dependent Variable: CLV

b. Selecting only cases for which Training = 1

$$y = f(Sat\_Mean\_1\_3, SMEAN(loans), SMEAN(long), SMEAN(age), SMEAN(income), (SMEAN(income))^2, (SMEAN(income))^4, e^{loy_{1_3}}, (SMEAN(loans))^2, (SMEAN(loans))^3, (SMEAN(loans))^5, sat1\_val3\_avg)$$

Figure.16: *Function of CLV and Model Coefficients*

Following on with this it is worth noticing that few variables do not lie in the conventional zone of significance, and two are highly insignificant with p-values > 0.8

Thus this model cannot be considered as an apt representation and thus is discarded, but this gives us many insights to what can be done later on.

Model B

This is simplified version of Model A. It is comparatively simpler and employs the WLS approach as well. However, R^2 is better but it violates certain known points, which shows that there is a chance to better this model further.

It is explained in further detail in Appendix E.

Model Summary[b,c,d]

Model	R Training = 1 (Selected)	R Training ~= 1 (Unselected)	R Square	Adjusted R Square	Std. Error of the Estimate
1	.429[a]	.460	.184	.173	24565.69219

a. Predictors: (Constant), sat1_val3_avg, SMEAN(loans), SMEAN(income), SMEAN (age), loy_1_3_exp, loans_power_3, loans_power2

b. Unless noted otherwise, statistics are based only on cases for which Training = 1.

c. Dependent Variable: CLV

d. Weighted Least Squares Regression - Weighted by SMEAN(long)

Figure.17: *Model summary for Model-B*

Model C

In this model to deal with the insignificance observed earlier for *loan* (and *loan-derived*) variables, we now include the WLS weights on the basis of an interaction variable of *savings* and *long*. This helps us in capturing the variance in loan as it is on the same scale as *long*.

The model summary is shown below:

Model Summary[b,c,d]

Model	R Training = 1 (Selected)	R Training ~= 1 (Unselected)	R Square	Adjusted R Square	Std. Error of the Estimate
1	.467[a]	.502	.218	.206	196076.2226

a. Predictors: (Constant), sat1_val3_avg, SMEAN(loans), SMEAN(age), SMEAN (income), loy_mea_12346, loans_power_5, loans_power2, loans_power_3

b. Unless noted otherwise, statistics are based only on cases for which Training = 1.

c. Dependent Variable: CLV

d. Weighted Least Squares Regression - Weighted by sav_long

Figure.18: *Model summary from the final model*

This model has the best accuracy from amongst all other models around **22%** of the change in CLV can be explained by this model. Also, it is imperative to say that the model has similar results for both the training and validation datasets, so it qualifies as a **robust model** and can be expected to give similar results in real world.

To analyse the variance in the data we read the **ANOVA table**, from this we can see that the F-stat is highly significant, which leads us to say that the model is good on this front as well.

ANOVA[a,b,c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.609E+12	8	7.011E+11	18.237	.000[d]
	Residual	2.007E+13	522	3.845E+10		
	Total	2.568E+13	530			

a. Dependent Variable: CLV

b. Weighted Least Squares Regression - Weighted by sav_long

c. Selecting only cases for which Training = 1

d. Predictors: (Constant), sat1_val3_avg, SMEAN(loans), SMEAN(age), SMEAN (income), loy_mea_12346, loans_power_5, loans_power2, loans_power_3

Figure.19: *ANOVA table for the final model*

The coefficients (Fig. 19) are all variables are significant and lie well within the critical value obtained from the t-test.

This model uses linear relation with *income*, which was not seen during the visual analysis. This was investigated and best results were found to be linear. (see Appendix C for related models).

$$y = f(SMEAN(income), SMEAN(age), SMEAN(loans), loans^2, loans^3, loans^5, loy\_mea\_12346, avg(sat1\_val3))$$

Coefficients[a,b,c]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-357.022	1799.760		-.198	.843
	SMEAN(income)	927.275	161.734	.233	5.733	.000
	SMEAN(age)	35.361	16.833	.082	2.101	.036
	SMEAN(loans)	321.337	137.939	1.780	2.330	.020
	loans_power2	-13.823	4.635	-8.857	-2.983	.003
	loans_power_3	.170	.049	11.603	3.446	.001
	loans_power_5	-5.895E-6	.000	-4.223	-3.622	.000
	loy_mea_12346	832.297	224.554	.164	3.706	.000
	sat1_val3_avg	-429.543	158.738	-.117	-2.706	.007

a. Dependent Variable: CLV

b. Weighted Least Squares Regression - Weighted by sav_long

c. Selecting only cases for which Training = 1

Figure.20: *All coefficients are significant in this model*

Conclusion

From the above analysis of data we can say that the CLV can be predicted by mixing variables and applying mathematical and statistical techniques.

The final suggested model can explain over 20% of CLV, and the variables having the most effect are found to be those of *income* and *loyalty-based* questions, which can be expected that loyal customers with more than average income will bring greater profitability overall.

Age and loans can be seen as high contributors to the total CLV also, and it is understandably so. A middle aged person will have contributed highly to the bank's profitability owing to higher income and length of relation. And someone who has loans with our bank will lead to higher profits due to interest. So, in its entirety the final model shows that it is a good estimator of the CLV and is in coherence with existing marketing theories.

One thing noticed in this project is that CLV is not highly dependent on the gender, education, profession of the customers and thus including them did not give a very high increase in R^2 but added to the model complexity.

Recommendations

A few survey recommendations that can be made are:

i. Ensuring that dirty data is not captured (such as gender=0).
ii. Missing values in the dataset lead to difficulty in analysis.
iii. Several responses were not shared as part of the dataset (*resp_no*). This should be avoided.
iv. Surveys should preferably be done on the same scale, 1-7 for loyalty and 1-10 for others.
v. Survey questions should be formed with utmost care, and they should all represent the same idea. For instance, Loyalty Q-5 was found to be entirely out of sync with the others, and Satisfaction Q-1 & Value Q-3 represent higher CLV far more than the others.

Some Marketing strategy recommendations are:

i. Income group specific strategies could be introduced, like with 0% credit cards for young people in lower income groups.
ii. Mortgage related schemes could be thought of for richer persons in middle age.
iii. Savings interests could be increased as persons with higher savings drives CLV.
iv. People aged between 43-48 years is the major customer base and provides the bulk of bank's profitability, so a focussed study on their needs is also recommended.
v. A study could be done to know what customers perceive as *value, satisfaction*, and *loyalty*, as the model shows that the Perception of benefits drives profitability.

References

Best, Roger J., 2012. To maximise long-term profits Available at: www.rogerjbest.com. (Accessed: 15th March 2019)

Deely, M, 2011. Profit Or Loss? The Difference Is Customer Lifetime Value. Available at: https://www.bigskyassociates.com/blog/profit-or-loss-the-difference-is-customer-lifetime-value. (Accessed: 05th March 2019)

IBM Corporation. Linear Regression. Available at: https://www.ibm.com/support/knowledgecenter/en/SSLVMB_sub/statistics_mainhelp_ddita/spss/base/idh_regs.html. (Accessed: 13th March 2019)

Jensen, K. CRISP-DM: Available at: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf (Figure 1), CC BY-SA 3.0, Accessed from: https://commons.wikimedia.org/w/index.php?curid=24930610/. (Accessed on: 17th March 2019)

Limam, M, 2012. Modelling customer lifetime value in retail banking context. Available at: https://www.academia.edu/27455201/Modelling_customer_lifetime_value_in_retail_banking_context. (Accessed: 17th March 2019)

R. Arboretti Giancristofaro, L. Salmaso, 2003. MODEL PERFORMANCE ANALYSIS AND MODEL VALIDATION IN LOGISTIC REGRESSION. Published in *Statistica*, anno LXIII, n. 2, 2003

Ryd´en, J, 2004. Multiple linear regression. Model validation, model choice. Available at: Department of Mathematics, Uppsala University.

Sean X. Chen and Jun S. Liu, Statistical Applications Of The Poisson-Binomial And Conditional Bernoulli Distributions, *Statistica Sinica,*Vol. 7, No. 4 (October 1997)

Shapiro. R. D. and Heskett. J. L. 1985. Logistics Strategy: Cases and Concepts. Universitas Michigan: Publisher West Pub. Co., 1985.

Skhmot, N., The 8 Wastes of Lean, (2017). Available at: https://theleanway.net/The-8-Wastes-of-Lean. (Accessed: 31st October 2018)

Appendices

Appendix A: Scatter Plots

Back to Methodology

CLV-Income

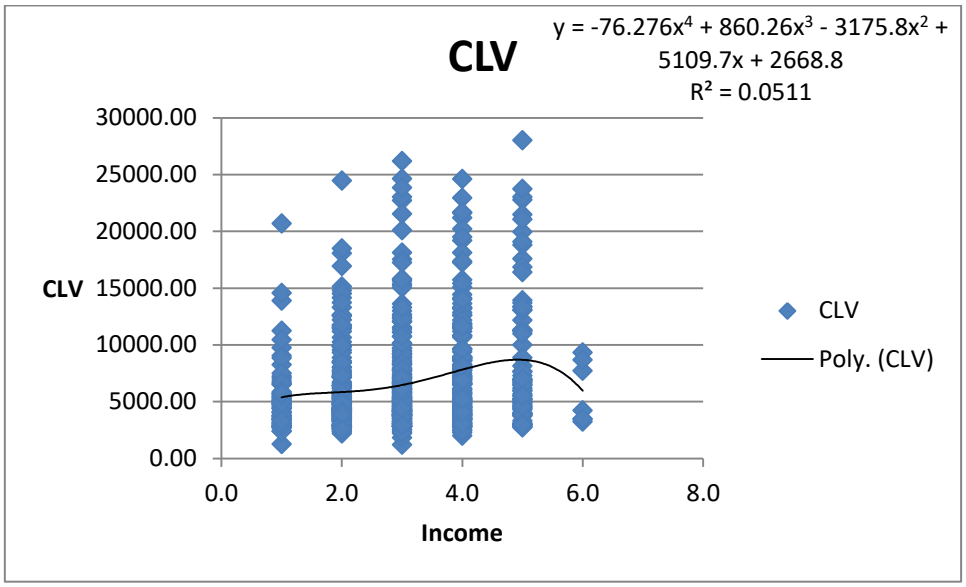

Figure.21: *Scatter plot between CLV v/s Income*

CLV-Loyalty1,2,3

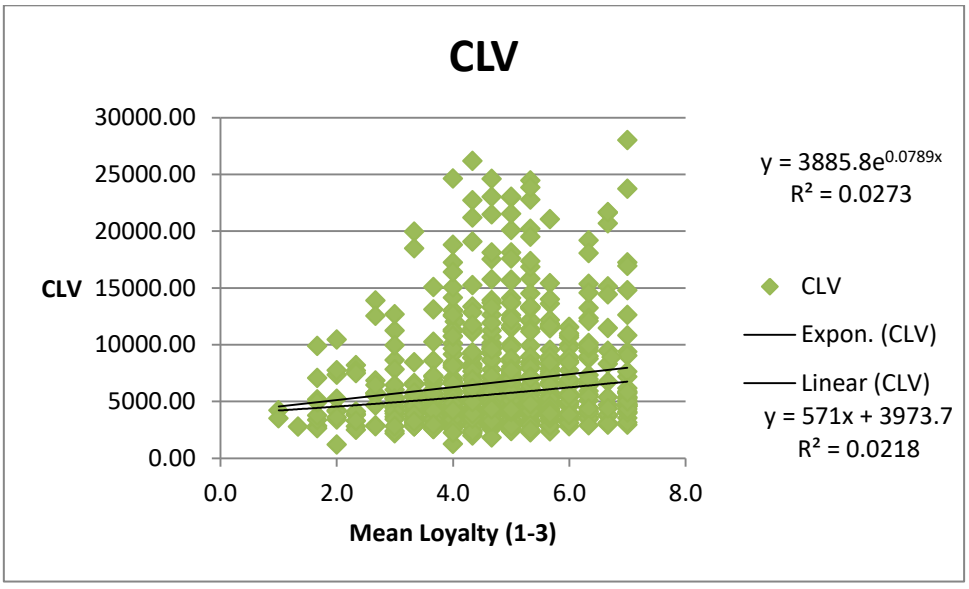

Figure.22: *Scatter plot between CLV v/s Loyalty_1-3*

Back to Methodology

Back to Methodology

CLV-Loyalty (1-6)

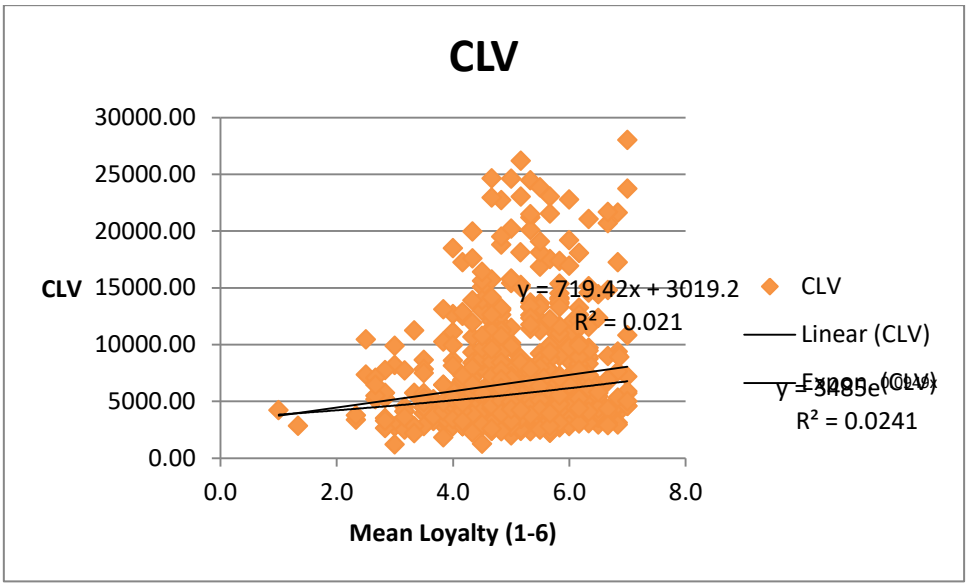

Figure.23: *Scatter plot between CLV v/s Loyalty_1-6*

CLV-Age

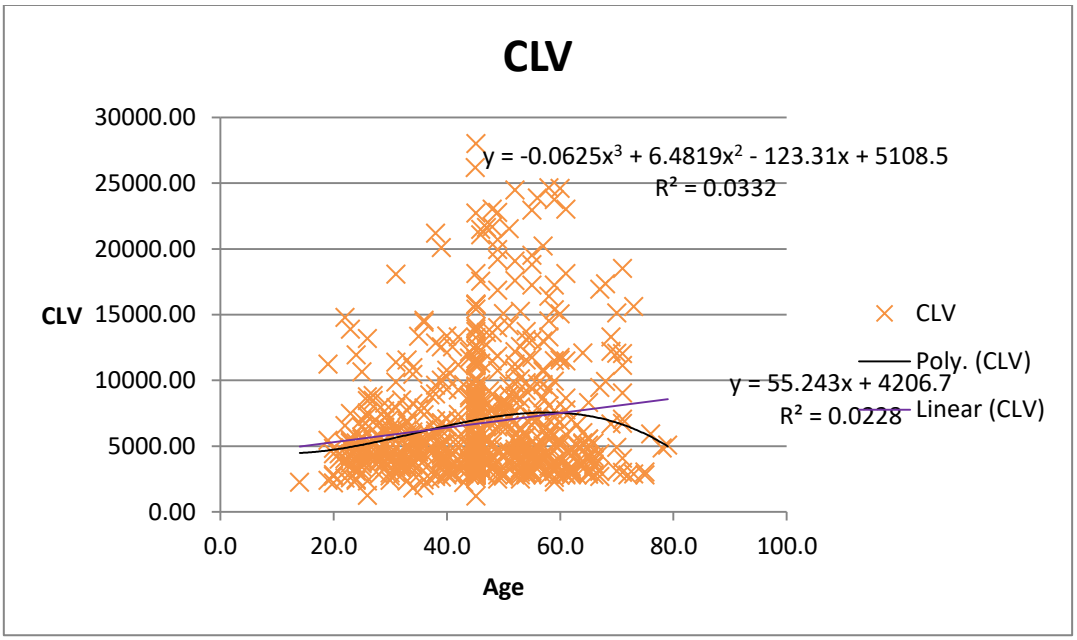

Figure.24: *Scatter plot between CLV v/s Age*

Back to Methodology

We can also see the effect of age on CLV, and the relation becomes very clear from the below scatter plot. The customers around 45 years provide the most benefit to the bank, and insights such as this could help in building our strategies. The data however follows a trend

which increases linearly as can be seen from the data. There are also two trend-lines which also show the R^2 values for these estimates. This analysis helps in building our model.

CLV-Loans

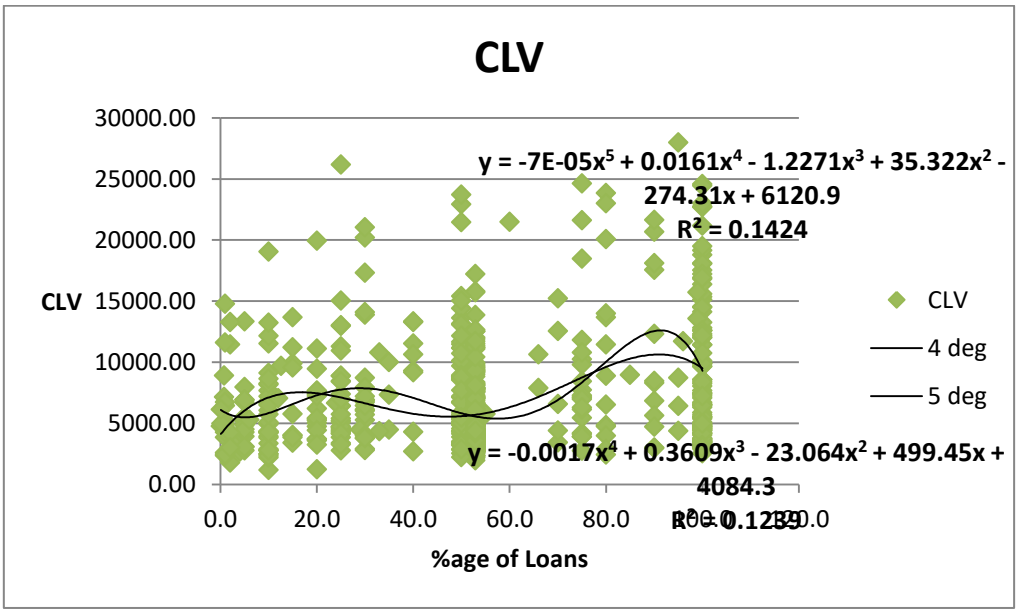

Figure.25: *Scatter plot between CLV v/s Loyalty_1-3*

CLV-Savings

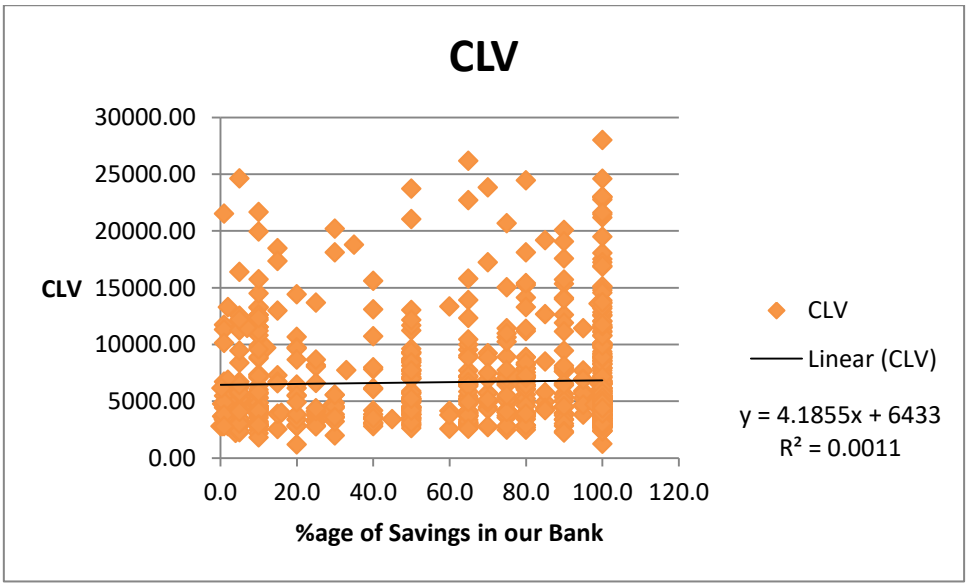

Figure.26: *Scatter plot between CLV v/s Savings*

Back to Methodology

CLV-Education

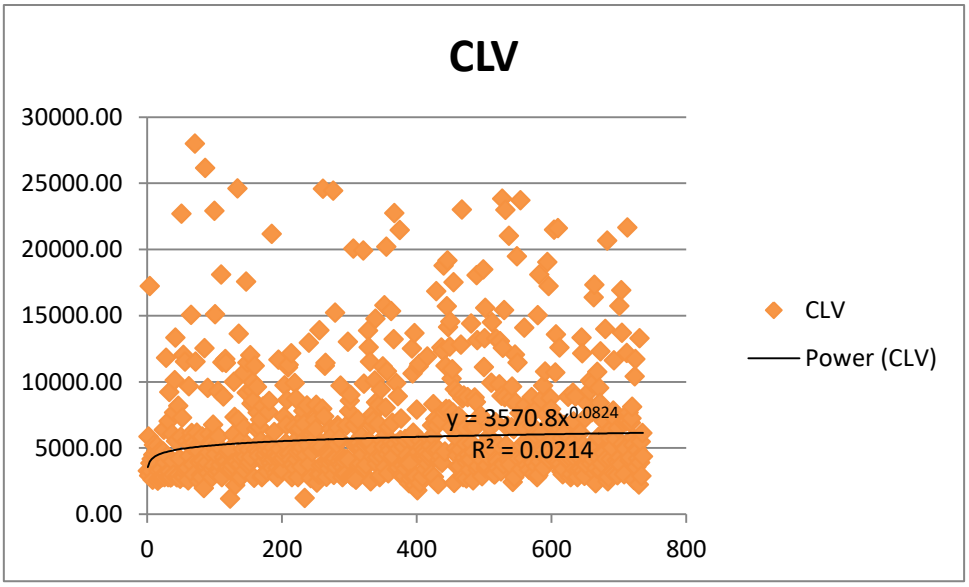

Figure.27: *Scatter plot between CLV v/s Education*

CLV-Branch

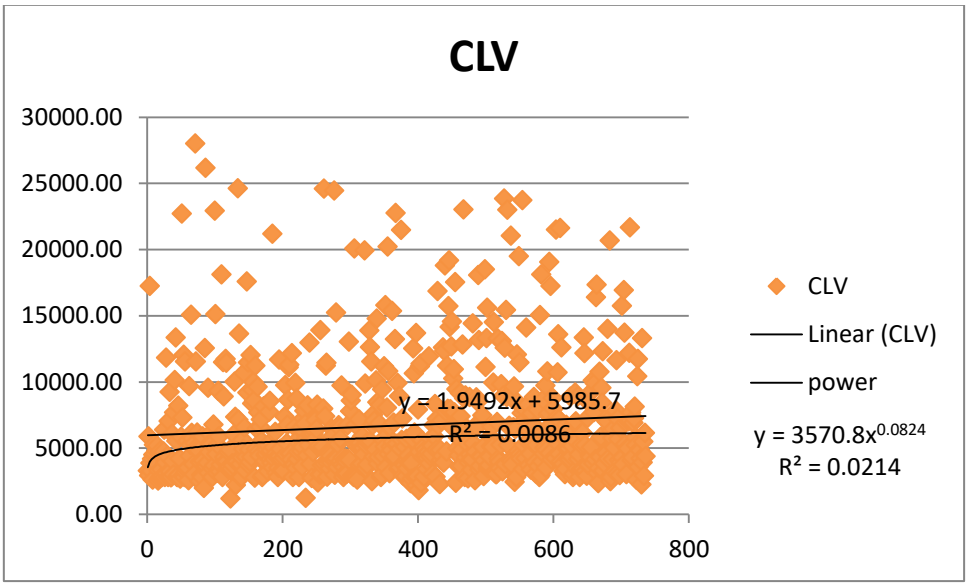

Figure.28: *Scatter plot between CLV v/s Branch*

Back to Methodology

Appendix B: Correlation Matrix

Correlations

			resp_no	branch	profess	gender	educatio	SMEAN(loy1)	SMEAN(loy2)	SMEAN(loy3)	SMEAN(loy4)	SMEAN(loy5)	SMEAN(loy6)	SMEAN(val1)	SMEAN(val2)	SMEAN(val3)	SMEAN(sat1)	SMEAN(sat2)	SMEAN(sat3)	SMEAN(sav)	SMEAN(loans)	SMEAN(long)	SMEAN(age)	SMEAN(income)	CLV	Loy_Mean_1_6	Val_Mean_1_3	Sat_Mean_1_3	Val_Mean_1_3_with2	Loy_Mean_1_3
Spearman	resp_no	Correlatio	1	0.003	0.064	0.014	-0.083	0.065	0.069	0.017	0.04	0.039	0.07	0.035	0.037	0.051	0.039	0.029	-0.009	0.016	0.032	-0.007	-0.098	0.022	0.137	0.07	0.05	0	0.049	0.065
	branch	Correlatio	0.003	1	-0.013	0.008	0.052	-0.048	-0.017	-0.005	0.011	-0.053	-0.052	-0.027	-0.041	-0.073	0.048	0.05	-0.018	0.052	-0.027	0.028	-0.004	0.026	0.022	-0.039	-0.046	0.016	-0.046	-0.017
	profess	Correlatio	0.064	-0.013	1	-0.065	-0.369	-0.133	-0.057	-0.052	0.05	-0.063	-0.072	-0.119	-0.156	-0.121	-0.07	-0.048	-0.087	0.023	0.032	0.032	-0.203	-0.085	-0.101	-0.085	-0.15	-0.081	-0.158	-0.107
	gender	Correlatio	0.014	0.008	-0.065	1	-0.12	0.129	0.083	0.035	0.033	0.081	0.094	0.051	0.051	0.084	0.08	0.056	0.094	0.117	0.041	-0.061	-0.015	-0.069	-0.015	0.116	0.071	0.082	0.068	0.098
	educatio	Correlatio	-0.083	0.052	-0.369	-0.12	1	0.013	-0.042	-0.075	-0.13	-0.07	-0.062	-0.008	0.025	-0.014	-0.068	-0.036	-0.055	-0.079	-0.006	0.04	0.208	0.131	0.029	-0.095	-0.003	-0.043	0.002	-0.046
	SMEAN(loy1)	Correlatio	0.065	-0.048	-0.133	0.129	0.013	1	0.385	0.283	0.198	0.405	0.547	0.397	0.42	0.52	0.325	0.308	0.384	0.127	0.029	0.073	0.112	0.036	0.107	0.65	0.482	0.372	0.474	0.627
	SMEAN(loy2)	Correlatio	0.069	-0.017	-0.057	0.083	-0.042	0.385	1	0.411	0.247	0.235	0.345	0.366	0.316	0.362	0.216	0.213	0.256	0.107	0.021	-0.002	0.174	0.017	0.144	0.702	0.378	0.256	0.37	0.812
	SMEAN(loy3)	Correlatio	0.017	-0.005	-0.052	0.035	-0.075	0.283	0.411	1	0.072	0.291	0.246	0.298	0.265	0.325	0.198	0.207	0.229	0.092	0.015	-0.044	0.059	-0.003	0.101	0.627	0.329	0.221	0.318	0.771
	SMEAN(loy4)	Correlatio	0.04	0.011	0.05	0.033	-0.13	0.198	0.247	0.072	1	0.175	0.298	0.158	0.19	0.224	0.13	0.112	0.128	0.077	-0.014	-0.066	-0.058	0.015	0.125	0.514	0.213	0.137	0.211	0.233
	SMEAN(loy5)	Correlatio	0.039	-0.053	-0.063	0.081	-0.07	0.405	0.235	0.291	0.175	1	0.603	0.412	0.422	0.501	0.332	0.293	0.364	0.076	-0.01	0.036	0.115	0.009	0.005	0.613	0.485	0.366	0.477	0.364
	SMEAN(loy6)	Correlatio	0.07	-0.052	-0.072	0.094	-0.062	0.547	0.345	0.246	0.298	0.603	1	0.515	0.53	0.636	0.401	0.367	0.452	0.07	-0.002	0.076	0.132	0.002	0.068	0.691	0.605	0.455	0.597	0.434
	SMEAN(val1)	Correlatio	0.035	-0.027	-0.119	0.051	-0.008	0.397	0.366	0.298	0.158	0.412	0.515	1	0.788	0.701	0.465	0.395	0.499	0.084	0.024	0.038	0.143	-0.096	0.055	0.499	0.903	0.506	0.885	0.434
	SMEAN(val2)	Correlatio	0.037	-0.041	-0.156	0.051	0.025	0.42	0.316	0.265	0.19	0.422	0.53	0.788	1	0.758	0.489	0.444	0.556	0.116	0.018	0.045	0.1	-0.045	0.057	0.505	0.907	0.56	0.942	0.411
	SMEAN(val3)	Correlatio	0.051	-0.073	-0.121	0.084	-0.014	0.52	0.362	0.325	0.224	0.501	0.636	0.701	0.758	1	0.57	0.488	0.62	0.082	0.02	0.084	0.128	-0.051	0.021	0.611	0.895	0.612	0.877	0.484
	SMEAN(sat1)	Correlatio	0.039	0.048	-0.07	0.08	-0.068	0.325	0.216	0.198	0.13	0.332	0.401	0.465	0.489	0.57	1	0.807	0.565	0.075	0.025	0.039	0.105	-0.036	0.016	0.387	0.539	0.801	0.537	0.313
	SMEAN(sat2)	Correlatio	0.029	0.05	-0.048	0.056	-0.036	0.308	0.213	0.207	0.112	0.293	0.367	0.395	0.444	0.488	0.807	1	0.616	0.051	-0.032	-0.009	0.106	-0.045	-0.018	0.357	0.463	0.847	0.466	0.304
	SMEAN(sat3)	Correlatio	-0.009	-0.018	-0.087	0.094	-0.055	0.384	0.256	0.229	0.128	0.364	0.452	0.499	0.556	0.62	0.565	0.616	1	0.065	-0.024	0.072	0.111	-0.077	0.041	0.425	0.596	0.89	0.597	0.35
	SMEAN(sav)	Correlatio	0.016	0.052	0.023	0.117	-0.079	0.127	0.107	0.092	0.077	0.076	0.07	0.084	0.116	0.082	0.075	0.051	0.065	1	0.13	-0.051	-0.175	-0.148	0.044	0.134	0.1	0.07	0.105	0.133
	SMEAN(loans)	Correlatio	0.032	-0.027	0.032	0.041	-0.006	0.029	0.021	0.015	-0.014	-0.01	-0.002	0.024	0.018	0.02	0.025	-0.032	-0.024	0.13	1	0.021	0.048	-0.091	0.091	0.004	0.029	-0.022	0.027	0.021
	SMEAN(long)	Correlatio	-0.007	0.028	0.032	-0.061	0.04	0.073	-0.002	-0.044	-0.066	0.036	0.076	0.038	0.045	0.084	0.039	-0.009	0.072	-0.051	0.021	1	0.093	0.086	0.011	-0.01	0.065	0.05	0.062	-0.011
	SMEAN(age)	Correlatio	-0.098	-0.004	-0.203	-0.015	0.208	0.112	0.174	0.059	-0.058	0.115	0.132	0.143	0.1	0.128	0.105	0.106	0.111	-0.175	0.048	0.093	1	0.159	0.105	0.121	0.135	0.127	0.132	0.15
	SMEAN(income)	Correlatio	0.022	0.026	-0.085	-0.069	0.131	0.036	0.017	-0.003	0.015	0.009	0.002	-0.096	-0.045	-0.051	-0.036	-0.045	-0.077	-0.148	-0.091	0.086	0.159	1	0.195	0.012	-0.063	-0.066	-0.058	0.018
	CLV	Correlatio	0.137	0.022	-0.101	-0.015	0.029	0.107	0.144	0.101	0.125	0.005	0.068	0.055	0.057	0.021	0.016	-0.018	0.041	0.044	0.091	0.011	0.105	0.195	1	0.142	0.065	0.021	0.067	0.158
	Loy_Mean_1_6	Correlatio	0.07	-0.039	-0.085	0.116	-0.095	0.65	0.702	0.627	0.514	0.613	0.691	0.499	0.505	0.611	0.387	0.357	0.425	0.134	0.004	-0.01	0.121	0.012	0.142	1	0.591	0.426	0.58	0.867
	Val_Mean_1_3	Correlatio	0.05	-0.046	-0.15	0.071	-0.003	0.482	0.378	0.329	0.213	0.485	0.605	0.903	0.907	0.895	0.539	0.463	0.596	0.1	0.029	0.065	0.135	-0.063	0.065	0.591	1	0.589	0.995	0.485
	Sat_Mean_1_3	Correlatio	0	0.016	-0.081	0.082	-0.043	0.372	0.256	0.221	0.137	0.366	0.455	0.506	0.56	0.612	0.801	0.847	0.89	0.07	-0.022	0.05	0.127	-0.066	0.021	0.426	0.589	1	0.592	0.348
	Val_Mean_1_3_with2	Correlatio	0.049	-0.046	-0.158	0.068	0.002	0.474	0.37	0.318	0.211	0.477	0.597	0.885	0.942	0.877	0.537	0.466	0.597	0.105	0.027	0.062	0.132	-0.058	0.067	0.58	0.995	0.592	1	0.475
	Loy_Mean_1_3	Correlatio	0.065	-0.017	-0.107	0.098	-0.046	0.627	0.812	0.771	0.233	0.364	0.434	0.434	0.411	0.484	0.313	0.304	0.35	0.133	0.021	-0.011	0.15	0.018	0.158	0.867	0.485	0.348	0.475	1

Figure.29: *Correlation Matrix*

[Back to Methodology]

Appendix C: Model C and higher powers of income and loan

Upon testing with the higher powers of income, it rendered both the one degree and higher degree terms as insignificant and thus dropped.

Coefficients[a,b,c]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	944.908	2020.323		.468	.640
	SMEAN(income)	-68.573	722.971	-.017	-.095	.924
	SMEAN(age)	33.695	16.858	.079	1.999	.046
	SMEAN(loans)	310.348	138.026	1.719	2.248	.025
	loans_power2	-13.511	4.636	-8.657	-2.915	.004
	loans_power_3	.167	.049	11.387	3.382	.001
	loans_power_5	-5.789E-6	.000	-4.147	-3.556	.000
	loy_mea_12346	824.274	224.412	.162	3.673	.000
	sat1_val3_avg	-402.474	159.739	-.110	-2.520	.012
	income_power2	163.339	115.582	.256	1.413	.158

a. Dependent Variable: CLV

b. Weighted Least Squares Regression - Weighted by sav_long

c. Selecting only cases for which Training = 1

Figure.30: *Coefficients after adding higher power to Model C*

Also, on employing the higher powers of *loan* the 4th degree was removed from the model as it was not able to properly justify the variation.

Back to Model C

Appendix D: **Loyalty based models**

(i) All loyalty points.

Coefficients[a,b]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	3285.028	1150.086		2.856	.004
	Loy_Mean_1_6	683.682	220.815	.133	3.096	.002

a. Dependent Variable: CLV

b. Selecting only cases for which Training = 1

Figure.31: *Coefficients for model on all Loyalty parameters*

Model Summary[b,c]

Model	R		R Square	Adjusted R Square	Std. Error of the Estimate
	Training = 1 (Selected)	Training ~= 1 (Unselected)			
1	.133[a]	.183	.018	.016	4620.21614

a. Predictors: (Constant), Loy_Mean_1_6

b. Unless noted otherwise, statistics are based only on cases for which Training = 1.

c. Dependent Variable: CLV

Figure.32: *Summary of model with all Loyalty parameters*

(ii) Using Loyalty 1,2,3 which have high correlation between them

Model Summary[b,c]

Model	R		R Square	Adjusted R Square	Std. Error of the Estimate
	Training = 1 (Selected)	Training ~= 1 (Unselected)			
1	.134[a]	.196	.018	.016	4619.84228

a. Predictors: (Constant), Loy_Mean_1_3

b. Unless noted otherwise, statistics are based only on cases for which Training = 1.

c. Dependent Variable: CLV

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4216.332	851.855		4.950	.000
	Loy_Mean_1_3	540.808	173.881	.134	3.110	.002

a. Dependent Variable: CLV

b. Selecting only cases for which Training = 1

Figure.33: *Coefficients and model summary for all Loyalty 1-3*

Back to Methodology

Appendix E: Model B

The key features of this model are:

(i) Lesser variables, thus lower complexity

(ii) WLS approach is applied, using variation of *long* for assigning weights

(iii) The R^2 value is equitable with that of the Model A and in fact the Adjusted R^2 (calculated on the basis of the number of variables) is higher.

(iv) All loan based variables are insignificant, which violates our initial visual analysis

We know that there is a higher power relation between loans and CLV, but it is not represented by this model. Thus, this gives us an idea that this model can be further honed and better results are feasible.

Model Summary[b,c,d]

Model	R Training = 1 (Selected)	R Training ~= 1 (Unselected)	R Square	Adjusted R Square	Std. Error of the Estimate
1	.429[a]	.460	.184	.173	24565.69219

a. Predictors: (Constant), sat1_val3_avg, SMEAN(loans), SMEAN(income), SMEAN (age), loy_1_3_exp, loans_power_3, loans_power2

b. Unless noted otherwise, statistics are based only on cases for which Training = 1.

c. Dependent Variable: CLV

d. Weighted Least Squares Regression - Weighted by SMEAN(long)

Figure.34: *Model summary for Model-B*

Back to Model B

So by reducing the number of variables and simplifying the model we are able to make better approximations in this model.

$$y = f(SMEAN(loans), SMEAN(age), SMEAN(income), e^{loy\_1\_3}, loans^2, loans^3, avg(sat1\_val3))$$

Coefficients[a,b,c]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-2587.922	2419.107		-1.070	.285
	SMEAN(loans)	-74.743	79.112	-.414	-.945	.345
	SMEAN(age)	28.439	16.415	.070	1.732	.084
	SMEAN(income)	995.750	158.374	.252	6.287	.000
	loy_1_3_exp	5216.645	1546.827	.149	3.372	.001
	loans_power2	1.077	1.878	.677	.573	.567
	loans_power_3	3.742E-5	.012	.002	.003	.997
	sat1_val3_avg	-318.241	152.631	-.092	-2.085	.038

a. Dependent Variable: CLV

b. Weighted Least Squares Regression - Weighted by SMEAN(long)

c. Selecting only cases for which Training = 1

Figure.35: *Equation and coefficients for Model-B*

Appendix F: IBM-SPSS Modeller Output File

Output1.spv

Link: https://drive.google.com/open?id=1PjBcM7mabNtF9PsVX_GbWLmlf7PJbpeL