2017

Pilgrim Bank: Customers Profitability

IGNACIO RECASENS

EXPLORATORY ANALYSIS USING R

Contents

Executive Summary

To determine who are the best customers for Pilgrim Bank in terms of profitability a random sample of atomic data at customer level was collected to analyze them by several variables such as Age, Income, Tenure years, home district, online capabilities usage and electronic Bill payment subscription. The main insight indicates that the bank should focus on migrating customers to electronic bill payments (currently 3% of customers and with profits 71% higher than offline) rather than focusing solely on migrating them to online banking. A model to predict profits was created, however the results show a low accuracy (MASE:0.86) since there's a high volatility of profits across variables. Further research is recommended to predict which customers should be targeted to persuade and migrate to electronic bill payment.

1. Introduction

This project consists of three parts; Data preparation where incorrect, incomplete or irrelevant records were cleaned and processed, Business Analysis to find insights that might help explain profitability of customers and finally a model selection to predict a customer's profit.

2. Data Preparation

The original sample dataset had 31.634 records and 8 variables, from which 33% of records presented missing data and had to be handled properly as discussed in the next section. The variable customer ID was irrelevant for the analysis and had to be removed from the data set. Furthermore, when performing predictive analytics metrics such as MPE and MAPE gave *"Inf"* values since some customers had profits equal to $0. These cases consisting of only 0.1% of the data were transformed to profits equal to $0.01 solving this issue with marginal impact on the model. Finally, 29 duplicated records and two inconsistent records with Electronic Bill Payments but not Online were removed.

2.1. Missing Data

The 33% of records that had some missing data were removed since the final sample without them still had more than 20 thousand records, enough to perform the required analysis without affecting it significantly. The main variables with missing data were Age and Income as shown in the table below:

Variable	Age	Income	Tenure	District	Online	BillPay	Profit	Total
Missing records	26%	26%	0%	0%	16%	16%	17%	33%

Table 1. Missing data

For some records where only one of the variables was missing, it was possible to predict the missing values, however they represented only 4.5% of the total data set. Because of the small amount of them and to avoid adding possible bias in the analysis they were removed. These cases were 466 records and 483 records were missing Incomes and Age respectively could have been predicted. There are other risk associated with inputting missing values that could lead to legal actions and/or mistaken insights.

2.2. Outliers

The remaining data had two clear outliers as Figure 1. shows below, a customer with a profit of $27,086 (17421% deviation from the mean) and another with -$5,643 (-3750% deviation from the mean), both removed since their cases are not representative of the data and would decrease the model accuracy.

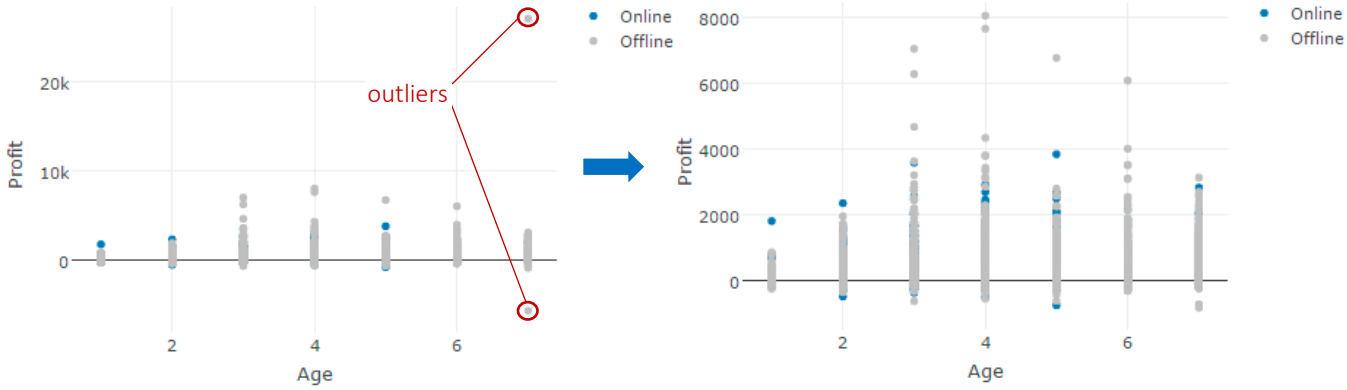

Figure 1. Outliers by profit

There were still some extreme points that could be taken care of according to the quartiles of the data as can be seen in table 2. were a capping technique was applied to smooth the extreme points. Every record before the 1% quartile and after the 99% quartile of profits was adjusted and fixed to the 1% and 99% quartile respectively.

Quartiles	0,1%	1%	10%	25%	40%	50%	75%	90%	99%	99.9%
Profit before	($305)	($182)	($78)	($29)	$0	$36	$226	$761	$1455	$3080
Profit afer capping	($182)	($182)	($78)	($29)	$0	$36	$226	$761	$1455	$1455

Table 2. Profits quantiles

3. Business Analysis

The average profit per customer is $146 but with a high standard deviation of $293. A summary of the main variables statistics can be found in Appendix 5.1.

3.1 Profit distribution

Even though the average profit per customer is $146 the bank is still losing a considerable amount of profits among its (worst) customers with negative profits that account for 40% of them (see table 2.) and with only 33% of customers with profits above the mean. This is also evident when looking at the median profit per customer which is $36, much lower than the average indicating that profit across customers is not evenly distributed.

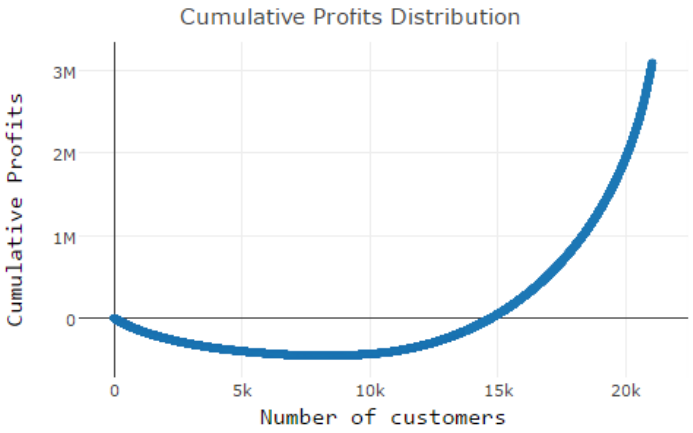

Figure 2. Cumulative profits

This is critical for the bank and they should penalize customers with negative profits charging them an extra fee to account for this. This can be done by determining the number of transactions required for Online and Offline customers and whenever they don't meet this quota, a fee should be charged to them.

There also seems to be a slight correlation of 0.2 between Income and Profit (see correlation matrix on Appendix 5.2) indicating that the highest the income the highest the profit (see Appendix 5.3).

3.2 Online/Offline channels penetration by Age

Unfortunately, the bank has still a long way to go in terms of online penetration, with only 20% of its customers under this format and even less under electronic bill payments with a mere 3% (see Table 3). As we'll see in the next section, this is not desirable and should be acted upon if profits are to increase.

	Bill Pay: Electronic	Bill Pay: Not Electronic	Total
Online	646 (3%)	3579 (17%)	4,225 (20%)
Offline	0 (0%)	16,825 (80%)	16,827 (80%)
Total	648 (3%)	20,404 (96%)	21,054 (100%)

Table 3. Online and electronic payment penetration by number of cases and % share over total.

As expected, there's a clear a clear affinity toward online channels among the younger generations where 82% of customers that use online channels are under 45 years old and 56% under 35 years old as can be seen through the Histograms in Appendix 5.4. This indicates that the younger the customer, the more comfortable she is to electronic banking. While it may be harder (i.e. too late) to convince old customers over 45 years old to change to online banking, implementing advertisements and incentives for younger generations might increase conversion rates and ROIs.

It can also be seen that inside the online market there's no difference on the demographic between customers subscribed to electronic payment as can be noted in Appendix 5.4, Figure A2.

3.3 Importance of Age and Online/Offline channels in Profits

Profits per customer varies across Age groups, especially over the extremes and less on the middle ranges as can be seen in Table 4. The highest profits are also perceived by customers using the online channels with a 15% higher profits per customer than offline. However, a distinction must be made between online customers with electronic bill payments and online customers without electronic bill payments. In fact, the main driver that explains this difference in profits is when customers are subscribed to electronic bill payments with a 71% increase in customer per profit vs. Offline as can be seen in Appendix 5.5, Figure A5.

Age group	All customers	Offline	Online (vs. Offline)	Online NOT Electronic BillPay (vs. Offline)	Online Electronic BillPay (vs. Offline)
< 15 years old	$ 33	$ 31	$ 37 (+17%)	$ 37 (+17%)	$ 33 (+6%)
15-24 years old	$ 89	$ 79	$ 110 (+38%)	$ 94 (+18%)	$ 199 (+150%)
25-34 years old	$ 147	$ 134	$ 188 (+41%)	$ 177 (+32%)	$ 243 (+82%)
35-44 years old	$ 157	$ 148	$ 195 (+32%)	$ 180 (+22%)	$ 274 (+86%)
45-54 years old	$ 153	$ 149	$ 178 (+19%)	$ 156 (+5%)	$ 331 (+123%)
55-64 years old	$ 166	$ 166	$ 173 (+4%)	$ 153 (-7%)	$ 278 (+68%)
>= 65 years old	$ 195	$ 196	$ 197 (+1%)	$ 196 (+0%)	$ 199 (+2%)
Total	**$147**	**$ 143**	**$ 163 (+15%)**	**$149 (+5%)**	**$ 245 (+71%)**

Table 4. Average Profit per Customer by channel

Since Profits are not normal (results validated by the Shapiro test and visually inspecting the data all in Appendix 5.6) to determine if the differences in profits were statistically significant the Wilcoxon test was required, although the t-test which assumes normality was also considered since there were more than 100 observations for each group. The results are presented below:

	Online (vs. Offline)	Online NOT Electronic BillPay (vs. Offline)	Online Electronic BillPay (vs. Offline)
Wilcoxon Test (p-value)	0.03723	0.5548	1.98e-13
t-test (p-value)	7.555e-05	0.241	4.249e-12

Table 7. Testing significance against Offline segment

From the results, even though it appeared that Online might have a significant difference with Offline profits, by separating though electronic bill payments it's clear that the real difference occurs when customers are subscribed to electronic bill payments. Even though Online might have higher profits than Offline, this difference is not statistically significant. It's clear then that the bank should focus on the main objective of migrating as much customers as possible to electronic Bill payments. One alternative is to

target the age segments that provide the highest profits as we saw before, the other is to target customers that are more likely to migrate (logit forecast).

4. Predictive Analysis

There's a clear correlation between Age and Tenure, however Income doesn't seem to be affected by Age or Tenure contrary to what common sense would have expected. There is however a slight correlation among all of them with profit; the higher the Age, Tenure and/or Income the higher the profit as can be seen through the graph in Appendix 5.8.

Tenure seems to correlate positively up to 15 years, however after this point there's a negative relationship. Including this as a piecewise variable "*Tenurep*" the adjusted r-square for *Profit ~ Tenure* to *Profit ~ Tenure + Tenurep* increase by 6% from 0.020 to 0.021. It was also included in the final model.

4.1 Model Selection

The first model to predict Profit considering all variables and observations reveals the previous findings regarding relationships and interactions. The model is as follows:

$$\mathbf{Profit} = 20.5 + 13.5 * \mathbf{Age} + 19.2 * \mathbf{Income} + 3.8 * \mathbf{Tenure} + 24.4 * \mathbf{District1200} + 15.5 * \mathbf{District1300}$$

$$-15.2 * \mathbf{Offline} - 88.9 * \mathbf{Not\_electronic\_Bill\_Pay} + \boldsymbol{\epsilon}\ ; Residual\ Error = 285.6,\ Adjusted\ R - squared = 0{,}053$$

All variables had p-values under 0.05 (for the exception of District1300 which had 0.07) suggesting that all variables are significant predicting profit, even demographics as can be seen in Appendix 5.7. As expected Age, Income and Tenure all had positive relationships with Profit being Income with the highest impact among the three where for each increase in Income Bucket the profit would increase by $19.2. As for categorical variables, District1200 provides the highest positive impact for profit followed by District1300 (District 1100 as reference with $0 increase), as for offline customers and/or not subscribed to electronic bill payment there was a negative relationship being BillPay with the highest impact where profits would be -$88.9 lower for customers without electronic payments.

Overall the model had a very low adjusted r-square explaining only 0.053 of the total variation and a standard error of the residuals of 285.6. When selecting a model evaluated over the training data (comprised of a sample of 75% total records) the best fit resulted in the following model:

$$\mathbf{Profit} = -74.4 + 27.2 * \mathbf{Age} + 13.24 * \mathbf{Income} + 3.33 * \mathbf{Income^2} - 2.92 * \mathbf{Income} * \mathbf{Age} - 2.26 * \mathbf{Income} * \mathbf{Offline} - 14.7 * \mathbf{Income} * \mathbf{Not\_Electronic\_BillPay} + 5.46 * \mathbf{Tenure} - 2.98 * \mathbf{Tenurep} + 23.9 * \mathbf{District1200} + 19.7 * \mathbf{District1300}\ ;\ Residual\ Error = 292.5,\ Adjusted\ R - squared = 0{,}060$$

Once again, we can see the same relationships as the previous model, however this time with a slight difference on their impact. The nonlinear quadratic relationship with Income indicates that the higher the Income the profits increase even more. However, the new interactions between Income*Age, Income*Offline and Income*Not_Electronic have negative relationships being the latest with the highest impact and in particular because customers with high incomes may produce higher transactions costs and in case of not being electronic it gives a higher negative impact than if they were electronic. This goes in line with the findings of acquiring profitable customers to join electronic payments. During the model selection process (Appendix 5.9) Model 5 was selected even though it had a lower r-squared compared to model 4 because metrics such as CV and AICc were lower due to the fact of having less variables which is desirable since we already have many variables being in higher risk of overfitting.

4.2 Model Prediction

The accuracies of the training and test data predicted through the previous models can be seen in Appendix 5.10. During the model selection process predicting the test data all models had a similar behavior as the linear model selection process, reducing their measures each time (for a few exceptions in RMSE). However, it gave higher assurance of selecting model 5 which was the only one that reduced the MASE during the process. Performing the Diebold-Mariano Test to determine if the model 5 has less error (is better) than model 1 and a Naive alternative the results for the p-values were 0.3157 and 0.000 respectively. This means that while Model 5 is indeed better than a Naive Alternative, it doesn't reveal a significant impact that would decide it's superiority over model 1 even though their accuracy metrics were better (but only slightly).

5. Conclusion

One of the main drivers for a higher customer profit besides Income is if they are subscribed to electronic payments which is a variable the Bank can act directly over customer to persuade them to migrate through incentives. Further research should be done to create a forecast model that predicts which customers are more likely to migrate and give a higher importance to those with a higher profit predicted from the forecast model stated on this document. While its accuracy wasn't high, it still serves as a starting point providing important insights of which variables affect the most.

6. Appendix

6.1 Main variables Statistics

Variable	Mean	Median	Standard deviation	Min	Max
Age Bucket	4,1	4.0	1.62	1	7
Income Bucket	5.5	6.0	2.33	1	9
Tenure	11.3	8.5	8.58	0.16	41.16
Profit	$146	$36	$293.4	($182)	$1.455

Table A1.
* Age buckets from 1 to 7 are Less than 15 years, 15-24 years, 25-34 years, 35-44 years, 45-54 years, 55-64 years and 65 years or older respectively.
* Income buckets from 1 to 9 are Less than $15k, $15k-$19k, $20k-$29k, $30k-$39k, $40k-$49k, $50k-$74k, $75k-$99k, $100k-$124k and $125k and more respectively.

6.2 Correlation Matrix

Figure A1. Correlation Matrix.

6.3 Profit vs. Income Boxplot

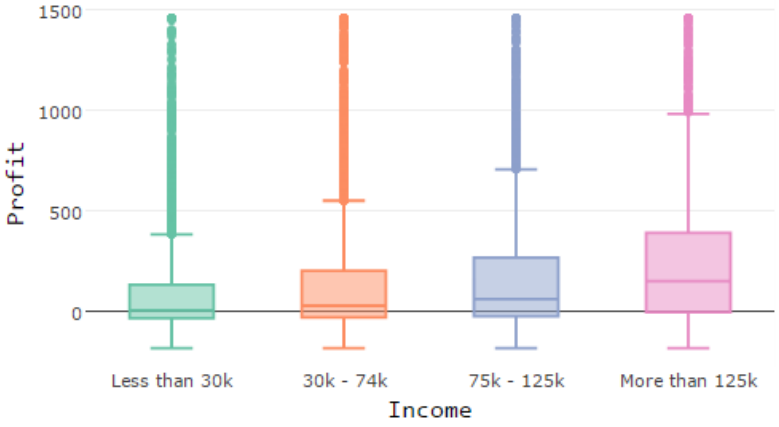

Figure A2. Income effect on Profits

6.4 Age distribution in Online/Offline market and Electronic Bill Payments subscriptions

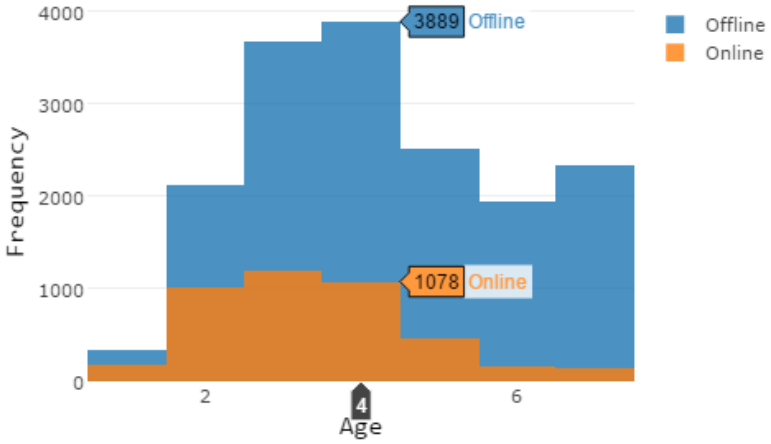

Figure A3. Age distribution in Online/Offline market

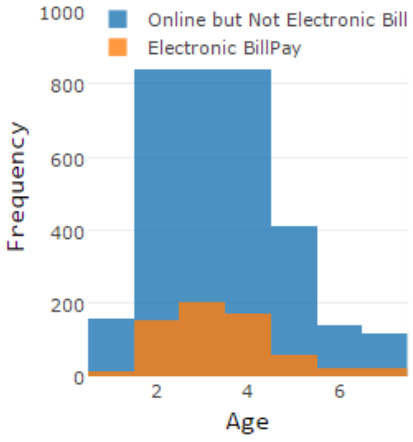

Figure A4. Age in Online market by Bill Payment

Age group	All	Online	Online without Electronic Payment %	Online with Electronic Payment %
Less than 15 years old	2%	4%	4%	1%
15-24 years old	15%	24%	24%	24%
25-34 years old	23%	28%	28%	32%
35-44 years old	24%	26%	25%	26%
45-54 years old	14%	11%	12%	9%
55-64 years old	10%	4%	4%	4%
65 years old or more	12%	3%	3%	4%
Total	100%	100%	100%	100%

Table A2. % Share over total by Age group.

6.5 Profit distribution in Online/Offline market and Electronic Bill Payments subscriptions

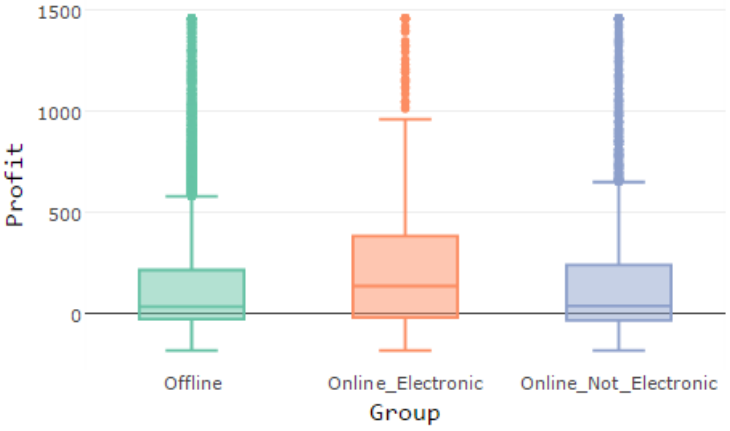

Figure A5. Profit by Online/Offline market and Electronic Bill Payments

6.6 Normality Test

Shapiro Test	Offline	Online	Online without Electronic Payment	Online with Electronic Payment
p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

Table A3. Normality test

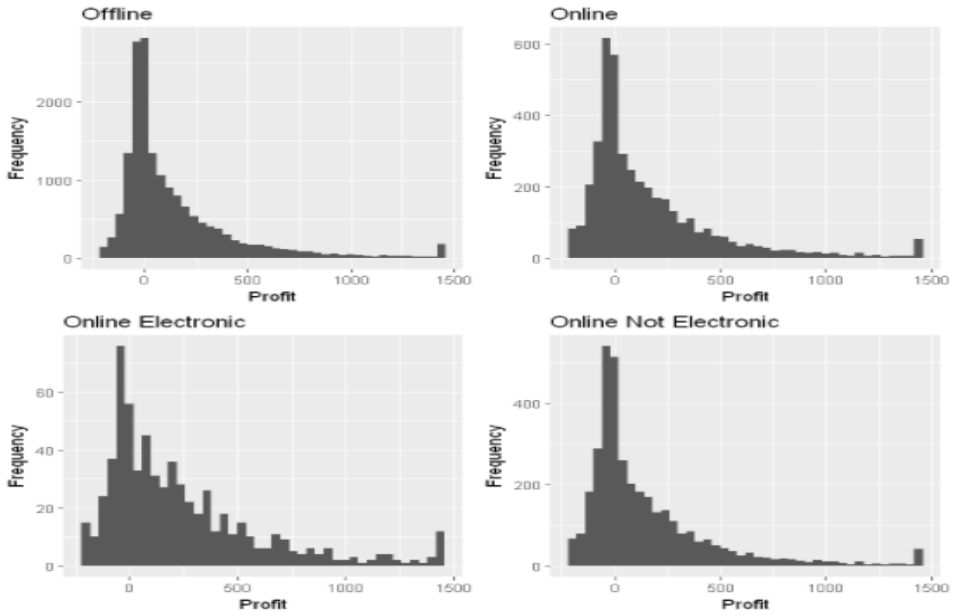

Figure A6. Profits histogram by Online/Offline channels and Electronic and not Electronic BIll Payments

6.7 Distribution across districts

District "1200" has the most profitable customers, mainly because they have the highest incomes.

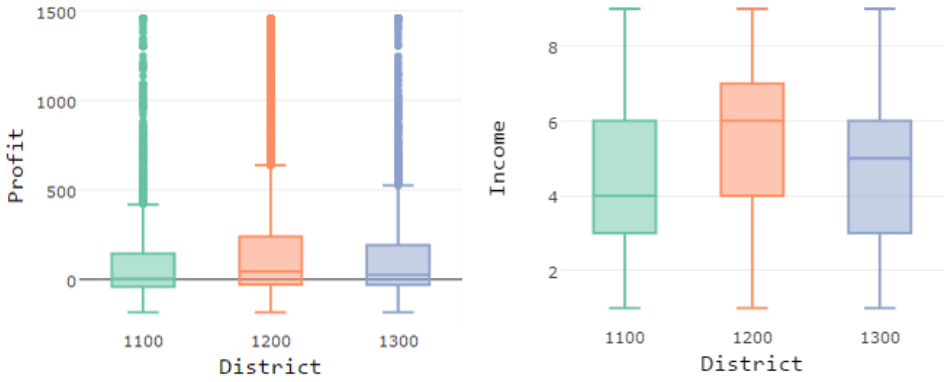

Figure A7. Profit and Income by District

6.8 Profit relationship with Income, Age and Tenure

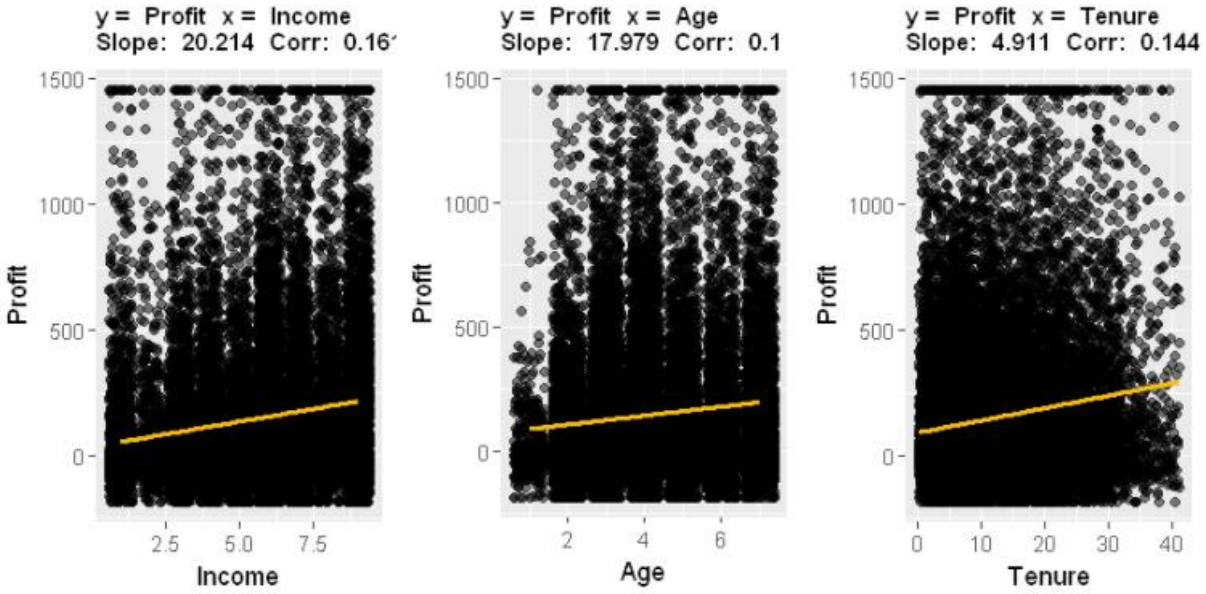

Figure A8. Profit vs Income, Age and Tenure

6.9 Linear Model selection process

Predictors / p-values	Fit 1	Fit 2	Fit 3	Fit 4	Fit 5
Age	0.000 ***	0.001 ***	0.000 ***	0.000 ***	0000 ***
Age^2	0.860	0.859	-	-	-
Income	0.013 *	0.012 *	0.012 *	0.012 *	0.129 *
Income^2	0.000 ***	0.000 ***	0.000 ***	0.000 ***	0.000 ***
Income * Age	0.000 ***	0.000 ***	0.000 ***	0.000 ***	0.000 ***
Income * Tenure	0.958	-	-	-	-
Income * Offline	0.341	0.342	0.339	0.024 *	0.025 *
Income * Not_Electronic_Bill	0.002 **	0.002 **	0.002 **	0.000 ***	0.000 ***
Tenure	0.000 ***	0.000 ***	0.000 ***	0.000 ***	0.000 ***
Tenurep	0.004 **	0.004 **	0.004 **	0.004 **	0.004 **
District1200	0.003 **	0.003 **	0.003 **	0.003 **	0.003 **
District1300	0.045 *	0.045 *	0.045 *	0.045 *	0.047 *
Offline	0.864	0.864	0.859	-	-
Not_Electronic_Bill	0.296	0.297	0.298	0.246	-
Adj. R-square	**0.06022**	**0.06028**	**0.06034**	**0.0604**	**0.06038**
CV	**86296**	**85653**	**85644**	**85635**	**85629**
BIC	**180634**	**180560**	**180550**	**180540**	**180532**
AICc	**180565**	**180445**	**180443**	**180441**	**180440**

Table A4. Linear Model Selection Process

6.10 Forecast Model selection process

Set	Model	Fit 1	Fit 2	Fit 3	Fit 4	Fit 5
Training Set	ME	-1.650e-14	-2.055e-14	2.360e-14	2.088e-15	3.450e-16
	RMSE	292.3862	292.3862	292.3865	292.3868	292.3992
	MAE	202.6247	202.6249	202.6170	202.6191	202.6045
	MAPE	551443.6	551507.8	551069.5	551259.6	550763.3
	MASE	0.9506671	0.9506681	0.9506307	0.9506407	0.9505722
Test Set	ME	-2.102e+01	-2.102e+01	-2.102e+01	-2.102e+01	-2.105e+01
	RMSE	258.6097	258.6121	258.6108	258.6111	258.5889
	MAE	185.1918	185.1908	185.1786	185.1769	185.1308
	MAPE	464604.7	464522.8	463906.5	463768.2	462242.0
	MASE	0.8688	0.8688	0.8688	0.8688	0.8685

Table A5. Forecast Model Selection Process