**Harvard Business Case Review:**

**Breaking Barriers: Micro-Mortgage Analytics**

*Written by:*

SUBRATA GARAI

*Submitted June 1st, 2017*

**INDEX**

|  |  |
| --- | --- |
| Introduction / Background | **3** |
| Problem Statement | **4** |
| Business Understanding | **4** |
| Data Understanding / Data Preparation | **7** |
| Interpretation of Logistic Regression Model 2 output (Exhibit 15) | **8** |
| Interpretation of CHAID Model Output (Exhibit 9 and 10) | **9** |
| Difference between Model 1 and Model 2 of Logistic regression | **11** |
| Trying to recreate the Logistic Regression model in R | **12** |
| Trying to recreate CHAID model in R | **13** |
| Validating the Models by the applicant's data | **14** |
| CHAID Model Prediction | **14** |
| Logistic Regression Model Prediction | **15** |
| Evaluation | **22** |

The case concerns Shubham Housing Development Finance Company (Shubham) in India. The company recognized the potential in serving the 20 million home-loan aspirants from low-income families (a market worth USD 182 billion in 2010). The basic challenge for Shubham is to evaluate the loan repayment ability of prospective clients in the absence of basic mortgage documentation used by regular banks. The chief operating officer, Ajay Oak, belief that the key to Shubham’s success is the quick identification of potential customers, fast application assessment, the ability to differentiate processing fees for creditworthy clients, and to ensure that all the processes are cost-effective.

Shubham start its lending operation in May 2011 with $2,000,000 (USD) in capital funding, opening a single branch in New Delhi. Within two years it had over 40 branches and raised an additional $7.8 million (USD) from venture capitalists. By September 2013, 2,300 applicants had received over INR 125 crore (USD 1=INR 62).

The loan evaluation process is an interview-based approach that assesses the applicants based on their daily or monthly cash earnings and expenses. Unorthodox compared to traditional mortgage assessment procedures; it required the presence of Shubham staff (credit officer, or CO) to verify the details provided by the applicants. The CO then creates a story about the applicant’s life, family, education, living conditions, income, liabilities, assets, work, and so on which forms the basis of their company profile. Ajay envisioned that an application scoring model could be build using the data generated from the field interactions to enhance decision making at the branch level. He believed that this would enable branches giving loans to better clients and the reduction of incorrect sanctions of more risky loan applications owing to subjective evaluations made by the local CO’s.

Based on the case description we assume that the desired top-line impact of evaluating prospective customers with the model application is the following:

1. Decreasing cost of loan originating about 10% versus the total transaction cost;

2. Potential to improve data intelligence; and,

3. Increased sales by amplified and more targeted marketing effort.

The bottom-line impact was a lowered overall cost of operation (especially due to less numbers of rejected loans), reduced competitive risk, and technological advantage over competitors.

**Problem statement:**

Our task is to understand the strengths and weakness of the analytics deployed by the Shubham Company micro-mortgage, to determine how they can create value using analytics - directly through cost reduction for loan origination and transaction, and indirectly through risk reduction and faster business realization for loan aspirants of low-income families.

**Business Understanding:**

The important issues of the case from the Management‘s perspective is the following:

1. **The management underestimate the value of its own analytics capability by not using any historical data:**

Shubham uses no data about existing customers and their behaviour patterns. For example; there is no repository of information indicating that paid their monthly fees regularly versus who dropped out on their payments / how many times, or why they dropped their payments.

Because there is no evidence from the existing data, the target variable definition of “the better client” or “incorrect sanction” are imprecise. For example, based on the previous period (2011-2013) how many clients of the 2,300 served are “good”? Why were those who were not good “bad”? What markers or features distinguished “bad” from “good”? The same is applicable to the definition faster business realization. Why is it taking so long and how much “faster” does Shubham want it to be?

1. **The technical and business teams are not working collaboratively before proceeding:**

A lack of communication within the analytics team can be proven by the use of different attributes to test classification techniques (e.g., a Junior Analyst used 21 attributes for logistic regression Model 1, whilst a Senior Analyst used 12 attributes for logistic regression Model 2).

There are also privacy and security concerns for the data collected by the CO of all interviewed people, regardless of if they are approved or rejected. No information is provided about the procedures of data acquisition and collection process at every stage of applicants’ assessment, or how it is stored after that process is complete. Shubham notes the data is “among its greatest sources of competitive advantage”, but there appears to be a lack of discipline surrounding it.

1. **The assumption that current staff have all necessary expertise and all business areas evaluated for possible impact is overstated:**

A flaw in the interview system employed by Shubham is the dependence on individual CO’s skills – some may be very good at determining quality customers, some may not be. Interviews create a potentially subjective evaluation system which can unduly influence decision-making.

Since the CO’s are the major source of the data acquisition, collection, and validation, they represent a large link in the chain which could go wrong, hurting the organization. Was the training provided for credit officer (CO) in interview techniques, data acquisition and collection up to the company’s standard? Who verifies, if anyone, that the data collected in the field is done so correctly? These sort of questions – not applicable to traditional lending institutions – are part-and-parcel of pathfinding in a new industry/market segment and need to be addressed.

There are a wide spectrum of socioeconomic backgrounds and demographic variation in customers, but they are currently all put in the same pool. The lack of external data for each customers’ geographic locations in finer-grain detail is also a potentially missed opportunity, even within the same city areas can have wildly different prices.

1. **No evaluation for increased risk of cost reduction or cost-benefit analysis is presented (only savings if using a model application). Thus, we have following questions to the management:**

How will cost reduction affect already the established loan evaluation process? How will cost reduction affect COs and the process of the evaluation of prospective customers?

The privacy, security, and regulatory risks of analytics are not evaluated in the model application. It is concerning that the customers ID consisted of the first three letters of the city they live in, and an alphanumeric string which could represent the person’s initials, street address, or postal code; as the full data is not available it cannot be said for sure. Randomized or hashed IDs with no PII would be the most secure option.

1. **No information is provided about how the current governance of the data will extend into the new system (or if it is required):**

What technology is required to support the COO’s vision? What is the investment needed (if any) into technology to handle the model application – a cost-benefit analysis here will be useful?

1. **Equality and missing data issues:**

The changes proposed need to be justified; why provide differential fee structure for differently sanctioned customers when insurance can be added to the customers who got fewer sanctions? How much will the property appreciated over the period of time, and why are some particular locations riskier that other – how will this be calculated?

There is no indexing which connects geographical location and population data (local area unemployment data, et cetera.)

**Data Understanding / Data Preparation:**

Logistic regression has the best performance at low probability of default thresholds. However, for higher thresholds CAIRD performs best in accuracy, precision, and specificity measures, while random forest performs best for negative predictive value and recall measures (M. Ali, (P. Andersson) 2015 “Performance of three classification techniques in classifying credit applications into good loans and bad loans: A comparison. Publisher Uppsala Universirsitet, 30p.)

# Interpretation of Logistic Regression Model 2 output (Exhibit 15):

Looking at the exhibit, it seems Logistic regression was applied based on different cut off values in an effort to find the most effective one. We derived the columns (Sensitivity, Specificity and Accuracy) from the given output data.

Where, Sensitivity (TPR) = TP / (TP + FN)

Specificity (TNR) = TN / (FP + TN)

Accuracy (ACC) = (TP + TN) / (TP + FN+ FP + TN)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cut-off | TP | TN | FP | FN |  | Sensitivity | Specificity | **Accuracy** |
| 0.05 | 1255 | 18 | 297 | 0 |  | 100 | 5.714286 | 81.0828 |
| 0.1 | 1255 | 22 | 293 | 0 |  | 100 | 6.984127 | 81.33758 |
| 0.15 | 1254 | 28 | 287 | 1 |  | 99.92032 | 8.888889 | 81.65605 |
| 0.2 | 1252 | 33 | 282 | 3 |  | 99.76096 | 10.47619 | 81.84713 |
| 0.25 | 1249 | 37 | 278 | 6 |  | 99.52191 | 11.74603 | 81.91083 |
| 0.3 | 1245 | 45 | 270 | 10 |  | 99.20319 | 14.28571 | 82.16561 |
| 0.35 | 1239 | 47 | 268 | 16 |  | 98.7251 | 14.92063 | 81.91083 |
| 0.4 | 1231 | 60 | 255 | 24 |  | 98.08765 | 19.04762 | 82.2293 |
| 0.45 | 1219 | 72 | 243 | 36 |  | 97.13147 | 22.85714 | 82.2293 |
| **0.5** | **1207** | **85** | **230** | **48** |  | **96.1753** | **26.98413** | **82.29299** |
| 0.55 | 1184 | 109 | 206 | 71 |  | 94.34263 | 34.60317 | 82.35669 |
| 0.6 | 1159 | 132 | 183 | 96 |  | 92.3506 | 41.90476 | 82.2293 |
| 0.65 | 1118 | 155 | 160 | 137 |  | 89.08367 | 49.20635 | 81.0828 |
| 0.7 | 1081 | 183 | 132 | 174 |  | 86.13546 | 58.09524 | 80.50955 |
| 0.75 | 998 | 217 | 98 | 257 |  | 79.52191 | 68.88889 | 77.38854 |
| 0.8 | 908 | 237 | 78 | 347 |  | 72.3506 | 75.2381 | 72.92994 |
| 0.85 | 772 | 264 | 51 | 483 |  | 61.51394 | 83.80952 | 65.98726 |
| 0.9 | 616 | 292 | 23 | 639 |  | 49.08367 | 92.69841 | 57.83439 |
| 0.95 | 391 | 307 | 8 | 864 |  | 31.15538 | 97.46032 | 44.4586 |
| 1 | 0 | 315 | 0 | 1255 |  | 0 | 100 | 20.06369 |

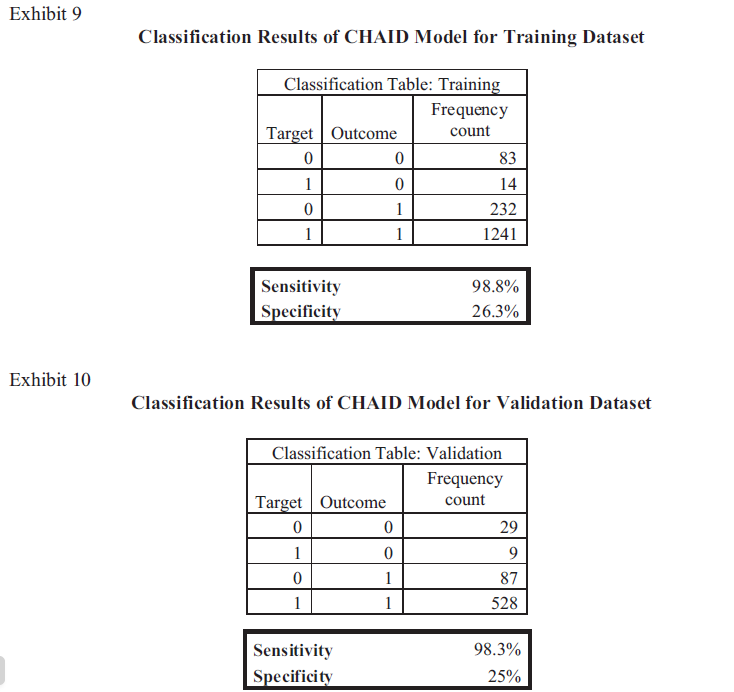
Table - Cut-Off versus Accuracy to was plotted to find the most effective cut-off value, highlighted in green.

Figure - Plot of Cut-off value versus Accuracy

This output data will be used to make a comparison between this model (Logistic Regression) and the alternate model (CHAID). The overall accuracy of this logistic regression model is a measure of the fit of the model. Here, this is .8229 which means that the model is estimated to give an accurate prediction 82% of the time.

# Interpretation of CHAID Model Output (Exhibit 9 and 10):

Exhibit 9 and 10 shows the output of this model for train and test set.



The table below converts the output data to the features found in Table 1 of the previous section (Sensitivity, Specificity and Accuracy)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP | TN | FP | FN |  | Sensitivity | Specificity | **Accuracy** |
| Testing | 1241 | 83 | 232 | 14 |  | 98.88446 | 26.34921 | 84.33121 |
| Training | 528 | 29 | 87 | 9 |  | 98.32402 | 25 | 85.29862 |

Table - CHAID model in terms of Sensitivity, Specificity, and Accuracy

The CHAID model has the greater accuracy (84.33%) as compared to the Logistic regression model (82.29%). Hence, using the CHAID model we have 84% chance of predicting the right candidate for approving or rejecting the loan.

# Difference between Model 1 and Model 2 of Logistic regression:

The models adopted by the Junior scientist (Keerthana) and the senior scientist (Siddharth) was the same - Logistic Regression. However, the difference is how the model was tested.

Looking at the output produced, Model 1was tested using ANOVA 'wald' test and Model 2 was tested using ANOVA 'ChiSq'.

***'ChiSq' Test – Model 2***

R Code for Logistic regression used was like below:

> model <- glm(Decision ~ Tier + Age + Yrsadd + Oldemi + FOIR + LTV, data=mmtrain, family=binomial(link = 'logit'))

Then, running *anova(my.mod, test="Chisq")*, the function compares the following models in sequential order:

glm(Decision ~1, family="binomial") **vs**. glm(Decision ~ Tier, family="binomial")

glm(Decision ~ Tier, family="binomial") **vs**. glm(Decision ~ Tier + Age, family="binomial")

glm(Decision ~ Tier + Age, family="binomial") **vs**. glm(Decision ~ Tier + Age + Yrsadd, family="binomial") and so on…

So, it sequentially compares the smaller model with the next more complex model by adding one variable in each step. Each of those comparisons is done via a likelihood ratio test (LR test; see example below).

***'Wald' Test – Model 1***

On the other hand, when we do Wald Test, the p-values in the output of summary(model) are Wald tests which test the following hypotheses (note that they're interchangeable and the order of the tests does not matter):

For coefficient of Tier: glm(Decision ~ Age +x3, family="binomial") **vs**. glm(Decision ~ Tier + Age + Yrsadd, family="binomial")

For coefficient of x2: glm(Decision ~ Tier +x3, family="binomial") **vs**. glm(Decision ~ Tier + Age + Yrsadd, family="binomial")

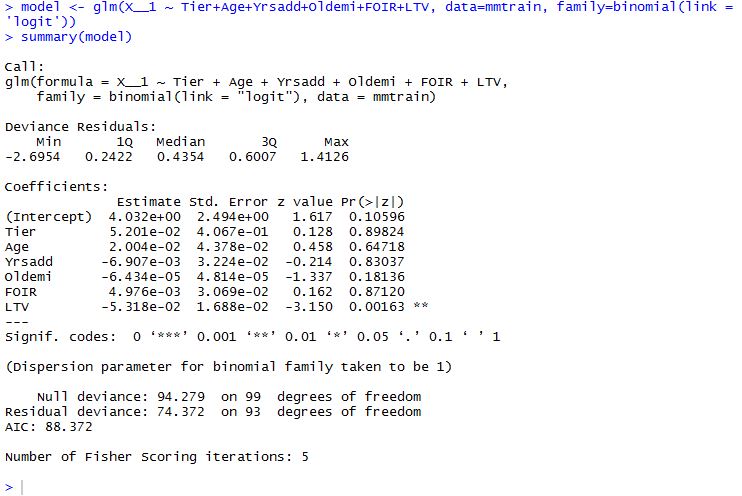
For coefficient of x3: glm(Decision ~ Tier +x2, family="binomial") **vs**. glm(Decision ~ Tier + Age + Yrsadd, family="binomial")

So, each coefficient against the full model containing all coefficients. Wald tests are an approximation of the likelihood ratio test.

# Trying to recreate the Logistic Regression model in R:

We tried to recreate the models using R to see how the output looks like compared to the output given.

Below is the model output in R for the 100 data points given:



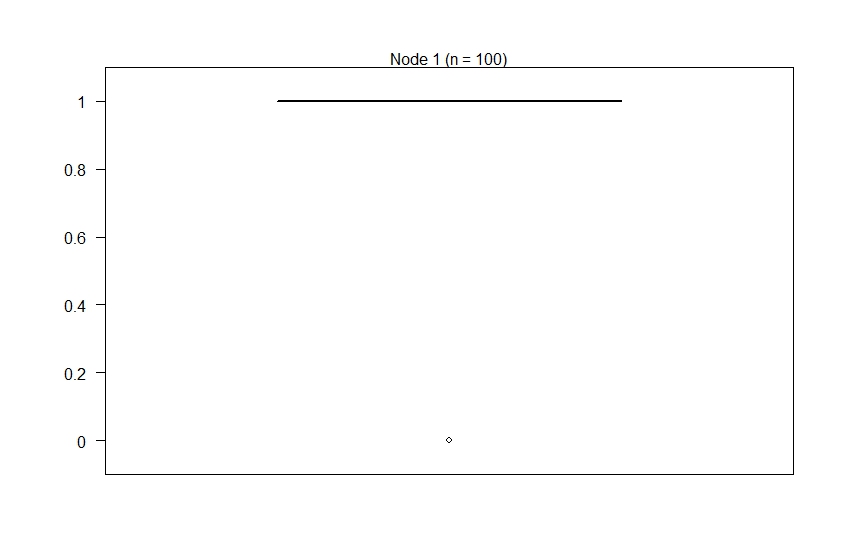
The output is significantly different (although we used z values) than the output provided. This can be explained by the fact that there are only 6 independent variables data is given for our use, whereas they used 20+ variables in their model using a complete dataset. Hence, the data given was not sufficient to recreate the results in Logistic Regression.

# Trying to recreate CHAID model in R:

General use of CHAID model is for large data sets.

When we tried to produce CHAID output using only 100 given data points, the output had no analytic value.

CHAID code



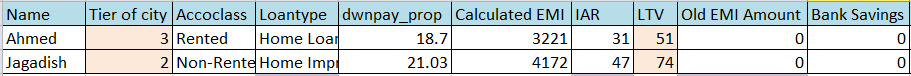
Hence, CHAID model couldn't be used for the little number of data points.

# Validating the Models by the applicant's data:

There are two prospective applicants mentioned in Exhibit 14. Some data points are mentioned to describe their socioeconomic status. We tried to validate these two cases through a CHAID decision tree and logistic regression.

**CHAID Model Prediction:**

The output of CHAID decision tree (Exhibit 8), uses three independent variables (LTV, FOIR and Tier of the City) to predict the dependent variable (Decision). We tried to find out these three variables from the data point given.



Two out of three variables are given – Tier of City and LTV

But, we need another variable (FOIR) in order to run these two applicants through the decision tree.

As per definition of variables given,

IAR = Calcemi / Monthly disposable income and

FOIR = (Calcemi + Oldemi)/Total monthly income

Dispinc (Disposable Income) = Total monthly income - Total monthly expenses

Considering the other variables given, we needed either 'Total monthly income' or 'Total monthly expenses' in order to calculate the FOIR and thereby taking help of CHAID model output. In the end, this was not available and given the presented exhibits, we were unable to predict their loan application through a CHAID model.

**Logistic Regression Model Prediction:**

We are providing tree examples for testing tree hypothesis.

***Question #1*:** Is approval or rejection applicant age depended?

***Hypothesis*:** Age of the applicants does not affect decision to provide loan.

Model 1-A (100 numbers of observations in the sample) is presented in Table 3 for the age of the applicants versus decision to accept or reject the loan application. A multiple regression close to 0 showed no correlation between age of the applicants versus decision to accept or reject the loan application. The R squared showed that only 3% fits the model. The standard error (7.8) is much higher that the coefficient of regression. ANOVA results showed that the t stat is 18.93 and significant (p-value<0.05), thus, the null hypothesis (age of the applicants does not affect decision to provide loan) is false.



Table - Model 1-A with age of the applicants versus decision to accept or reject the loan application

Figure - Residual plot presented for age of the applicants versus decision to accept or reject the loan application and showed that most of the data are not randomly dispersed around the horizontal axis, a non-linear regression model is appropriate for the data. Visually the decision to provide the loan (or not) are not depended on the age of the applicant because predicted age looks quite close between accept (1) and reject (0).

***Question #2*:** How does LTV (and market value) relate to the city tier (rank)?

***Hypothesis:*** LTV of the applicant’s property are the same in all city tier.

Model 1-B (100 numbers of observations in the sample) presented in Table 4 for the LTV (total loan requested/market value) of the applicants versus city tier where the loan was sought. A multiple regression of 0.14 showed no correlation between LTV of the applicants versus city tier. The R squared showed that only 22% fits the model. The standard error (0.76) is higher that the coefficient of regression. ANOVA results showed that the t stat is 11.71 and significant (p-value<0.05), thus, the null hypothesis (LTV of the applicant’s property are the same in all city tier) is false.



Table - Model 1-B with LTV of the applicants versus city tier where the loan was sought

Figure - Residual plot presented for LTV of the applicants versus city tear where the loan was sought showed that most of the data are for each of tree tier data are not randomly dispersed around the horizontal axis, and predicted numbers are closely aligned t to the data, however the trend is not very clear. Visually the more loans requested in tier#2 and LTV are the most different in city tier #1.

***Question #3:*** How affordable is the property for the people who applying to the loan?

***Hypothesis:*** LTV (market value) of the applicants’ property does not depend on FOIR.

Model 1-C (100 numbers of observations in the sample) presented in Table 5 for LTV (total loan requested/market value) of the applicants versus FOIR (CalcEMI (calculated monthly expenses) + EMI for earlier loan that applicant pays every month). A multiple regression close to 0 showed no correlation between age of the applicants versus decision to accept or reject the loan application. The R squared showed that only 6% fits the model. The standard error (10.3) is much higher that the coefficient of regression. ANOVA results showed that the t stat is 14.15 and significant (p-value<0.05), thus, the null hypothesis (LTV of the applicants’ property are the same as (or does not depend on) FOIR) is false.



Table - Model 1-C with LTV of the applicants versus FOIR

Figure - Residual plot presented for LTV of the applicants versus FOIR showed that data are randomly dispersed around the horizontal axis, and predicted numbers are closely aligned to the data, a linear regression model is appropriate for the data.

Multiple regression maybe used in addition to other models to analyze particular hypothesis, however limited. For example; as we showed the approval for loan correlated to age, the same age can be predicted to be in both approved and rejected groups. LTV (market value), monthly expenses of the applicant, and the geographical location (city tier) of the acquiring property were also interrelated.

**Evaluation:**

The case study illustrates the framework of the variables chosen and the ratios calculated, as well as the use of decision trees and logistical regression model.

As pointed out in the Business Understanding section, it is not evident from the case study on why the variables were chosen and what significance the ones that were chosen had over the ones that were not. It does not provide any clear indication on how these variables have business value.

The case study also fails to throw light on how the models were trained and the outcome of the hypothesis. We assume that they would have done that using the historical data about the customers who came in as first-time customers and the information about their failures or success in repaying loans which can be analysed on historical data. With the advent of external open/purchased data readily available in the market, it adds a new dimension towards predictive analytics.

This factor by and large is seen as missing. For example, Credit History Data (TransUnion from Canada perspective) holds a lot of insights into the applicant history from a financial laws perspective and regulations of the land. It provides an avenue to identify potential risk of the applicant being defaulted on any other mortgages, revolving credits, taxes, any recoveries et cetera.

The Affordability Index and House Price Index data provided by Statistics Canada also throws light on the overall market analysis of the applicant prospective property. The Domain experts can contribute significantly to use this data for segmentation and rating of the applicant.

Unemployment rate data provided by Statistics Canada also contributes significantly on which fields of employment are under strain and can give significant insight into analytics from the risk of job loss perspective.