Paper 11060-2016

Debt Collection through SAS® Analytics Lens

Karush Jaggi, AFS Acceptance; Thomas Waldschmidt, SquareTwo Financial;
Dr. Goutam Chakraborty, Oklahoma State University

ABSTRACT

Debt Collection! The two words can trigger multiple images in one's mind – mostly harsh. However, let's try and think positively for a moment. In 2013, over $55 billion of past due debt was recovered by agencies in the United States. What if all of these debts were left as is and the fate of credit issuers in the hands of good will payments made by defaulters? Well, not the most sustainable model to say the least. In this situation, debt collection comes in as a tool that is employed at multiple levels of recovery to keep the credit flowing. Ranging from in-house to third party to individual collection efforts, this industry is huge and plays an important role in keeping the engine of commerce running.

In the recent past, with financial markets recovering and banks selling less of charged off accounts and at higher prices, debt collection has increasingly become a game of efficient operations backed by solid analytics. This paper takes you in to the back alleys of all the data that is in there and gives an overview of some ways modeling can be used to impact the collection strategy and outcome. SAS® tools such as Enterprise Miner™ and Enterprise Guide™ are extensively utilized for both data manipulation and modeling. Decision trees are given more focus to understand what factors make the most impact.

Along the way, this paper also gives an idea of how analytics teams today are slowly trying to get the 'buy-in' from other stake holders in any company which surprisingly is one of the most challenging aspects of our job.

INTRODUCTION

While collection agencies would love to recover all the money owed from its customers in one go, it is a long road before any money is actually collected. Each customer account generally goes through a certain lifecycle that can be summarized in the below diagram.

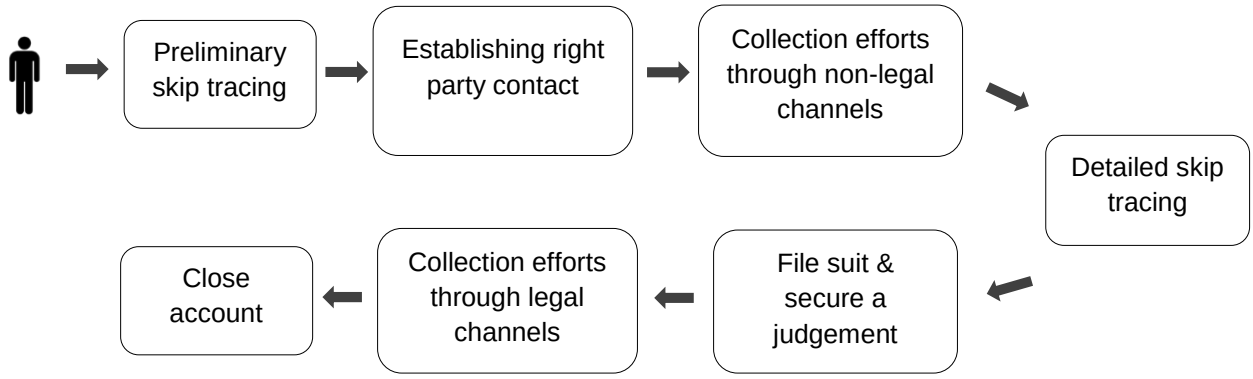


Figure 1. Debt collection lifecycle

In the figure 1, we see that once a customer is acquired, they are put through a journey that looks broadly the same for most of them barring some exceptions like customers above the age of 65 years who cannot be sued.

Beginning with preliminary skip tracing, every customer is put through a process of establishing certain basic details like existing phone number, address and credit report. The aim here is to get a general history of the customer that serves as a good starting point for making initial decisions about contacting them. Once their identity is mapped, the process of establishing contact with them is begun. It may come as a surprise, but even after all the skip tracing efforts, getting the right person on the phone or reaching them via snail mail is one of the biggest challenges affecting ROI. The next step after establishing a right party contact (RPC) is to begin non -legal collection efforts (unless they dispute the debt, which leads to another story). This step includes offering payment plans, settlement options, upfront payments etc. However, it is observed that a considerable chunk of customers are still unable to make the required payments after which their account proceeds to a more stringent legal channel. This channel involves a detailed skip tracing effort aimed at exploring details like properties owned by the customer or their employment details. This channel usually leads to a judgement giving the collections company options such a putting a lien on property or garnering wages. An important thing to note here is that the legal channel is an expensive option for the collections company considering it involves an array of court costs and attorney fees.

What is interesting to note here is that at all stages of this process, considerable amount of data collected and stored about each and every customer that makes way for statistical analysis and predictive modelling in a way that optimizes operational efficiency. It is not only in-house data that is available, but also a whole spectrum of third party data that can be bought off the shelf to aid such advanced analytics.

Getting in to the analytics side of this, a desired analytics process is one that will equip the collections agency with insights into basic questions like “What are the chances this customer will pay?”, “When is the right time to collect from this customer?”, “How much should be collected from this customer?” etc. With this in mind, the broad aim of this paper is really to understand what factors play an important role in maximizing collections and how we can capitalize on them. A few of the industry standard models that are usually used in tandem today are:

1. To predict if a customer is going to pay or not in its entire lifetime (Probability of Payment or PoP)
2. If yes, to predict the approximate month of first payment
3. To predict the $ amount that will be recovered from the customer

The results from all the above models will go into deciding the course of treatments an account will be subject to in its recovery lifecycle. For example, an account predicted to be a payer as well as predicted to make the first payment in month 6 from date of buying is a good indication to make collection efforts through traditional calling etc. However, an account predicted to be a payer but predicted to make the first payment in month 24 from date of buying is a sure shot indication to take legal action against them considering we know that all payers after month 18 are known to pay after being sued.

HOW TO GET PROBABILITY OF PAYMENT?

Most industry experts will concur that predicting the probability of payment is one of the most critical set of information to have while buying & working charged off accounts. This probability can be calculated for payments within a certain time frame (like 6 months) or over the life of an account. We will see ahead how to adjust the target variable to suit different needs. Naturally, predicting this probability is a relatively complex task that is affected by a number of external factors that are not captured by the system. However, a wide variety of internally available data is a good starting point for us.

To begin with, let us look at a few meaningful attributes about a customer that are available in-house.

ATTRIBUTE	TYPE	USABLE TRANSFORMATION
ACCOUNT_OPEN_DATE	Date	DAYS_SINCE_ACCOUNT_OPENED
CHARGEOFF_DATE	Date	DAYS_SINCE_CHARGEOFF
DELINQUENCY_DATE	Date	DAYS_SINCE_DELINQUENT
LAST_PREPURCHASE_PAYMENT_DATE	Date	DAYS_SINCE_LAST_PREPURCHASE_PAYMENT
FIRST_PAYMENT_MONTH	Date	FIRST_PAYMENT_MONTH_FROM_PURCHASE
CHARGEOFF_AMOUNT	Numeric	USE AS IS
ORIGINAL_PURCHASE_AMOUNT	Numeric	USE AS IS
ORIGINAL_LOAN_AMOUNT	Numeric	USE AS IS
ORIGINAL_CREDIT_LIMIT_AMOUNT	Numeric	USE AS IS
INTEREST_RATE	Numeric	USE AS IS

Table 1. In house customer attributes

As we can see, the date type variables are converted to a more usable format of number of days since the date captured. This can be easily performed using the SAS function INTCK (interval, from, to) as shown below.

```
DAYS_SINCE_DELINQUENT=
   INTCK("DAYS",DATEPART(LAST_PREPURCHASE_PAYMENT_DATE),
   DATEPART(PURCHASE_DATE)) + 1;
```

DATEPART function is used to extract date from a date time variable.

Moving forward, it is worth while discussing some of the third party data available today without much hassle. To begin with, the US Census provides a broad range of ZIP code level demographic data that can be crucial considering we know where our customers live. Some examples are median household income, median home value, median per capita income and more. These are generally good indicators of the neighbourhood that a customer lives in.

Other sources of third party data include credit bureaus like FICO®, Experian® and TransUnion® or public record aggregators like LexisNexis® and TLo. Let's look at some attributes that we can get from them.

ATTRIBUTE	RANGE/TYPE	SOURCE	DEFINITION
PROPENSITY RECOVERY SCORE 3.0	350-850	TransUnion®	Likelihood of collecting $50 or more within 12 months
YIELD RECOVERY SCORE 3.0	000-999	TransUnion®	Identifies accounts likely to pay more money
RECOVERY BANKCARD SCORE	350-850	TransUnion®	Likelihood of collecting $100 or more within 6 months
INCOME ESTIMATOR 3.0	000-999	TransUnion®	Estimates an individual's income based on credit history
FICO CREDIT SCORE	300-850	FICO®	Customer's credit score
VerifiedSSN	BINARY	LexisNexis®	Indicates if the customer's input SSN is verified
VerifiedPhone	BINARY	LexisNexis®	Indicates if the input phone is verified
VerifiedAddress	BINARY	LexisNexis®	Indicates if the input address is verified
InputAddrAgeNewestRecord	0-960	LexisNexis®	Indicates if the input address in address history
FelonyAge	0-84	LexisNexis®	Months since most recently recorded felony conviction

Table 2. Third party customer attributes

The above attributes only scratch the surface and are nowhere close to the exhaustive list that is available for deployment. We can easily access detailed data revolving around education, income, professional associations, address etc.

At this point, we are equipped with a lot of data that will be helpful in predicting probability of payment (PoP). To understand PoP, let's visualize a set of customers as show below.

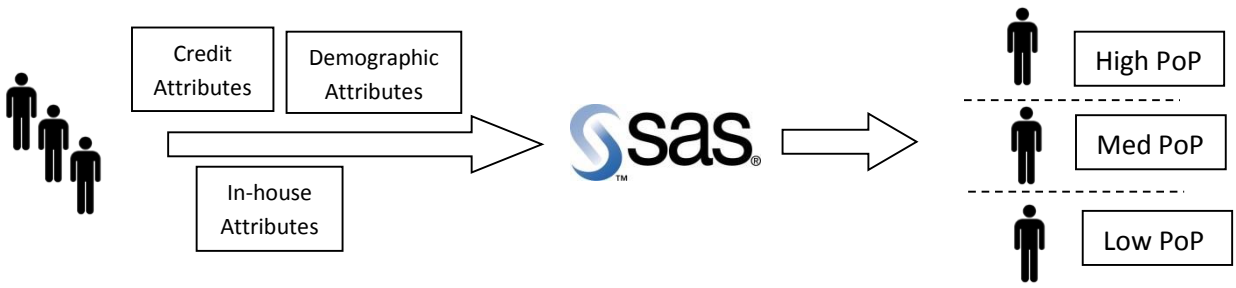


Figure 2. Visualizing Probability of Payment

In figure 2, we see that all of the customer attributes are used as input to SAS tools for modeling. To begin with, the first step here is to get insights into what attributes are actually useful and have some predictive value.

While there are multiple ways to do this, the Kolmogorov Smirnov statistics & charts are widely used in risk analytics. In its simplest form, this is basically a non-parametric test of equality of two probability distributions. Basically checking a null hypothesis that the two distributions are identical. If the KS statistic is large enough, the two are significantly different. Visually, this is measured as the distance between distributions of a positive outcome and negative outcome for the analysis variable.

Let's see how to do this.

We can use PROC NPAR1WAY to get KS statistics as shown below.

```
ODS GRAPHICS ON;

/*GENERATING KS STATISTICS & CHARTS*/

PROC NPAR1WAY EDF DATA=LIB.DATASET;
CLASS PAYER_FLAG ;
VAR ADDRCHANGECOUNT01 CTR3P2E;
OUTPUT OUT = LIB.OUTPUT (KEEP = _VAR_ _KS_ _D_);
RUN;
```

As can be seen, I have used the `EDF` option in the **`PROC NPAR1WAY`** statement that gives us KS statistic. A host of other options can be used to get other statistics.

The `CLASS` statement above can be used to specify the attribute we are trying to classify, payers versus non-payers in this case (*PAYER_FLAG*). `VAR` statement is optional and can be excluded if all variables are to be used.

`ODS GRAPHICS ON` ensures we get the below empirical distribution graph (figure 3).

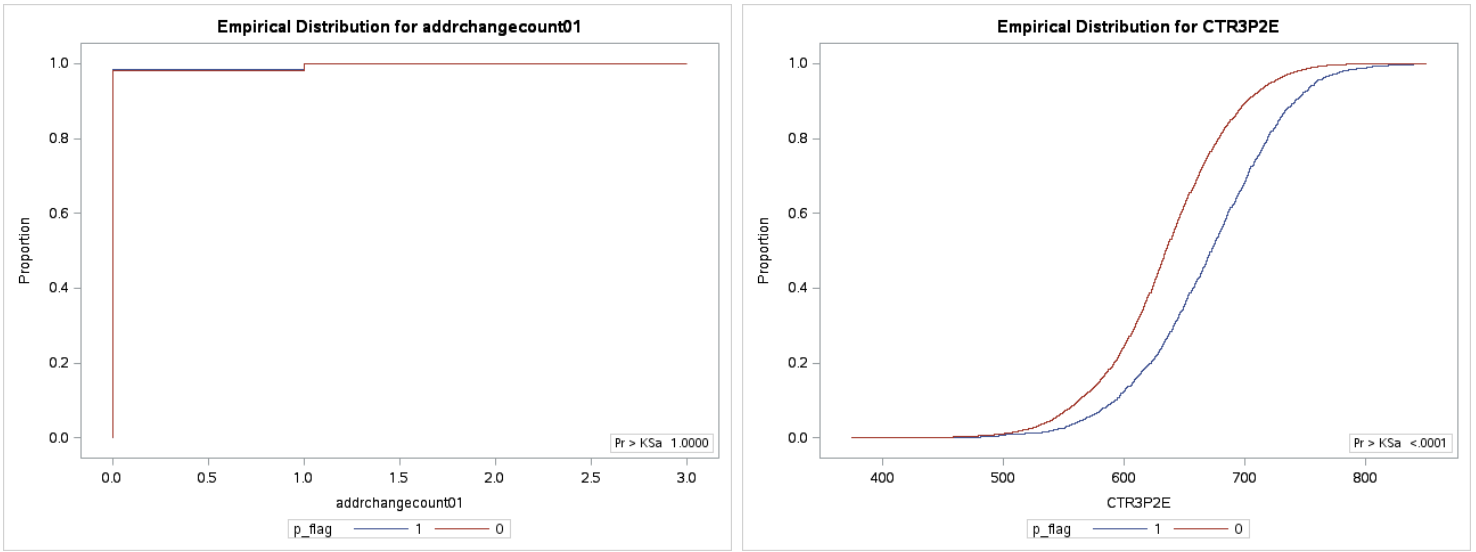


Figure 3. Empirical Distribution using KS Statistic

As can be seen above in figure 3, variable ADDRCHANGECOUNT01 is not very helpful in separating payers from non-payers. An important consideration with the KS Statistic is that since it is a univariate analysis, we do not get any other details such as collinearity of variables and hence we use multivariate analysis as described ahead.

This information about the predictive capability of an attribute is very crucial for third party data since most of them come at a certain price. When dealing with millions of customer records, this makes a considerable difference in the operational costs of collections.

Once we have identified the important variables, we can move to our next step which involves using a decision tree to see how the selected variables stack up and look at them from a business point of view.

Let us begin with this diagram from Enterprise Miner

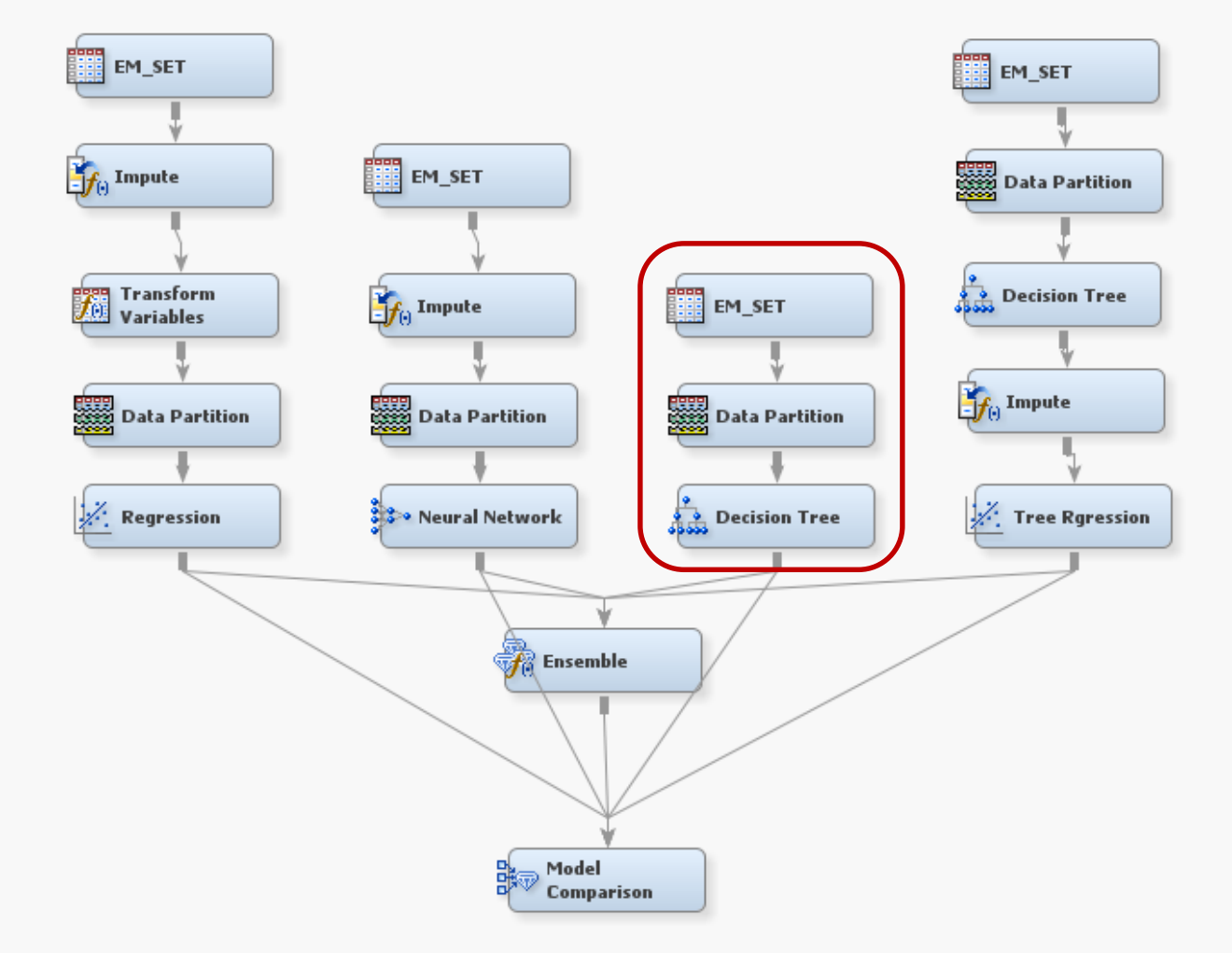
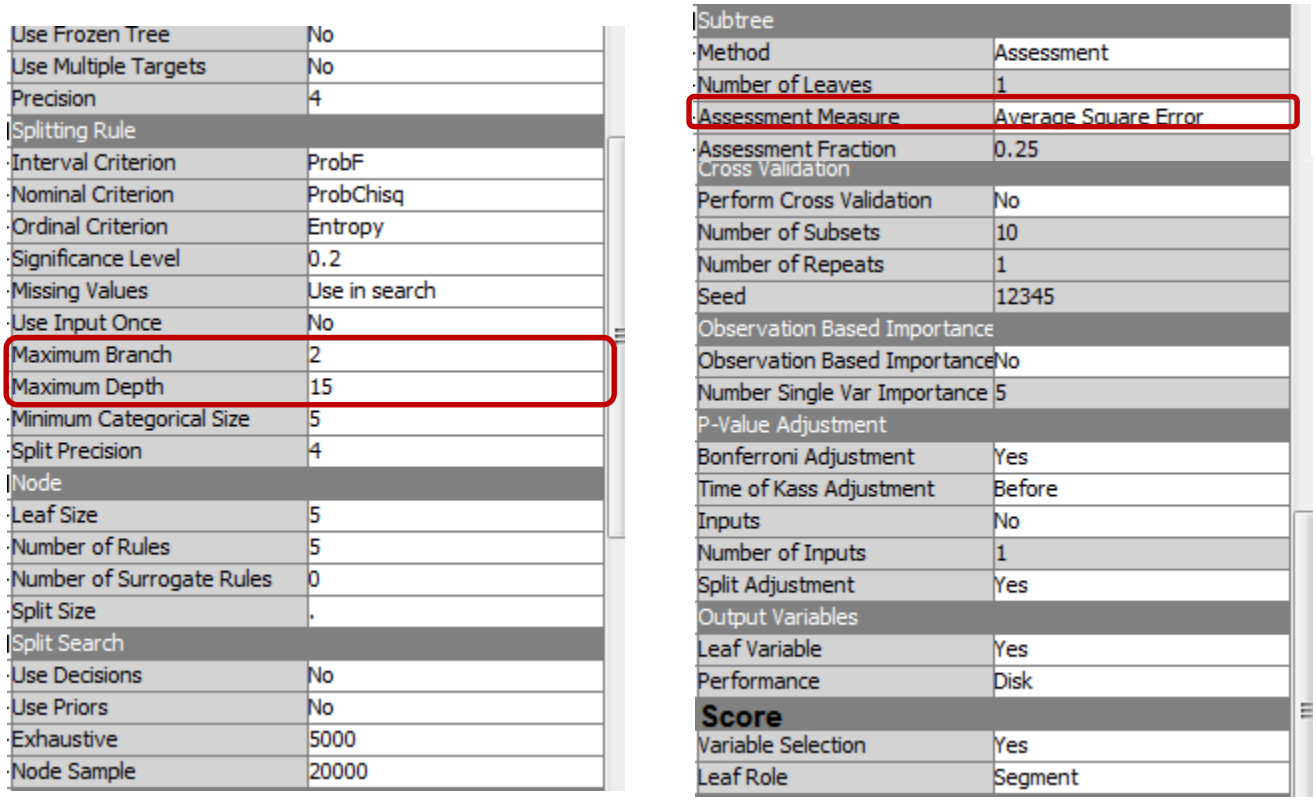


Figure 4. SAS® Enterprise Miner Diagram

In the above figure 4, we can see our data set EM_SET deployed with a simple regression, neural network, decision tree and tree regression.

Going into the decision tree, we set it up for rank ordering customers based on their propensity to pay and hence the target is the variable *PAYER_FLAG.* The detailed properties are as below in figure 5:



Use Frozen Tree	No
Use Multiple Targets	No
Precision	4
Splitting Rule	
Interval Criterion	ProbF
Nominal Criterion	ProbChisq
Ordinal Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	15
Minimum Categorical Size	5
Split Precision	4
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000

Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importance	
Observation Based Importance	No
Number Single Var Importance	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Kass Adjustment	Before
Inputs	No
Number of Inputs	1
Split Adjustment	Yes
Output Variables	
Leaf Variable	Yes
Performance	Disk
Score	
Variable Selection	Yes
Leaf Role	Segment

Figure 4. SAS® Enterprise Miner Diagram Properties

An important point to note here is the assessment measure chosen for the stopping point of the tree is average square error. We chose to use this average square error because the end goal of this model is to provide a rank order of accounts for incoming debt purchases. This rank order will be used in establishing operational strategies and prioritizing work effort. Also, please note the other options like max branch, max depth etc. are good to work with for optimizing complexity of the decision tree.

Decision tree have a variety of outputs that we can use. From estimating variable importance to scoring a new data set, they can be used for a variety of purposes. One of the more used application of decision trees is with regression as a variable selection technique.

Looking at the decision tree map show below, we get a very compact overview of the separation between the cases and non-cases. The node width is representative of the number of cases contained in the node after the split. Also, by default, the color of the node is reflective of the number of cases in the node. In figure 5 below, we can see that most of our cases (black nodes) are in the bottom right of the map which gives a sense of the payment pattern that we are trying to unravel from the data.

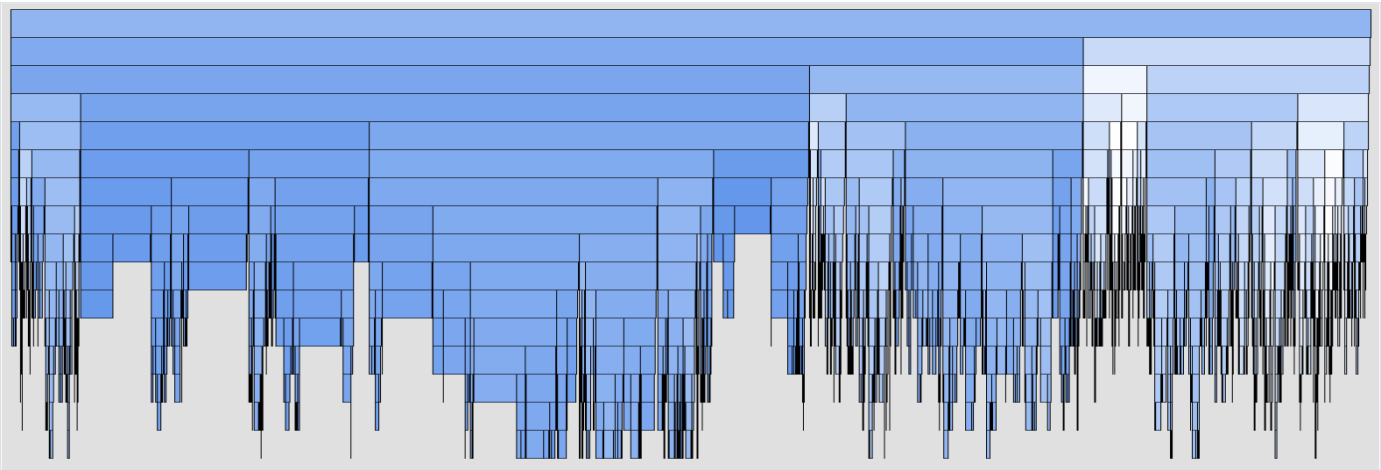


Figure 5. SAS® Decision Tree Map

Coming to the variable importance report generated by the decision tree, we have the following variables that are ranked by their predictive value. The importance measure below is basically a relative score given to each variable when compared to the most important variable, TU_RECOVERY_SCORE in this case. An interesting application of this score is to identify irregularities or inconsistencies with the node splits that the tree comes up with. This is done by comparing the importance of a variable with the vimportance (validation importance) for the same variable and if they are markedly different, we know there is an issue. The ratio column on the extreme right is a ratio of vimportance/importance that can also be used. A ratio of 1 is desired.

Variable Importance

Obs	NAME	LABEL	NRULES	IMPORTANCE	VIMPORTANCE	RATIO
1	TU_RECOVERY_SCORE	TU_RECOVERY_SCORE	128	1.0000	1.0000	1.0000
2	PRE_PRCH_PMT_TO_PURCHASE_DAYS	PRE_PRCH_PMT_TO_PURCHASE_DAYS	104	0.4921	0.4909	0.9977
3	ACTV_HM_CL_PHON_AT_PRCH_COUNT	ACTV_HM_CL_PHON_AT_PRCH_COUNT	34	0.2922	0.2979	1.0196
4	PURCHASE_STAGE	PURCHASE_STAGE	81	0.2224	0.2197	0.9881
5	PURCHASE_STATE_CODE	PURCHASE_STATE_CODE	79	0.2169	0.2059	0.9492
6	MINOR_DEBT_TYPE	MINOR_DEBT_TYPE	52	0.1806	0.1796	0.9942
7	PRE_PRCH_PMT_TO_CHARGEOFF_DAYS	PRE_PRCH_PMT_TO_CHARGEOFF_DAYS	79	0.1788	0.1758	0.9833
8	ORIGINAL_PURCHASE_AMOUNT	ORIGINAL_PURCHASE_AMOUNT	131	0.1778	0.1661	0.9341
9	SUBPRODUCT_TYPE	SUBPRODUCT_TYPE	113	0.1564	0.1517	0.9701
10	INTEREST_RATE	INTEREST_RATE	88	0.1326	0.1269	0.9571
11	PRODUCT_TYPE	PRODUCT_TYPE	30	0.1110	0.1074	0.9680
12	CHARGEOFF_TO_PURCHASE_DAYS	CHARGEOFF_TO_PURCHASE_DAYS	66	0.1106	0.0983	0.8895
13	DBTR_SEEN_BFR_PRCH_COUNT	DBTR_SEEN_BFR_PRCH_COUNT	68	0.1040	0.1011	0.9727
14	ACCT_OPEN_TO_PRE_PRCH_PMT_DAYS	ACCT_OPEN_TO_PRE_PRCH_PMT_DAYS	41	0.1000	0.0908	0.9079
15	DBTR_PAID_BFR_PRCH_COUNT	DBTR_PAID_BFR_PRCH_COUNT	68	0.0850	0.0856	1.0069
16	ACCT_OPEN_TO_CHARGEOFF_DAYS	ACCT_OPEN_TO_CHARGEOFF_DAYS	26	0.0645	0.0602	0.9326
17	ACTV_ADDR_AT_PRCH_COUNT	ACTV_ADDR_AT_PRCH_COUNT	28	0.0596	0.0472	0.7918
18	ACTV_WORK_PHONE_AT_PRCH_COUNT	ACTV_WORK_PHONE_AT_PRCH_COUNT	26	0.0471	0.0397	0.8412
19	ACCT_OPEN_TO_PURCHASE_DAYS	ACCT_OPEN_TO_PURCHASE_DAYS	9	0.0269	0.0181	0.6713

Figure 6. Decision Tree Variable Importance

Now that we have used the decision tree and have some results from it, we will deploy this decision tree in tandem with a regression model (also called a tree regression). Consider the highlighted branch of the diagram below.

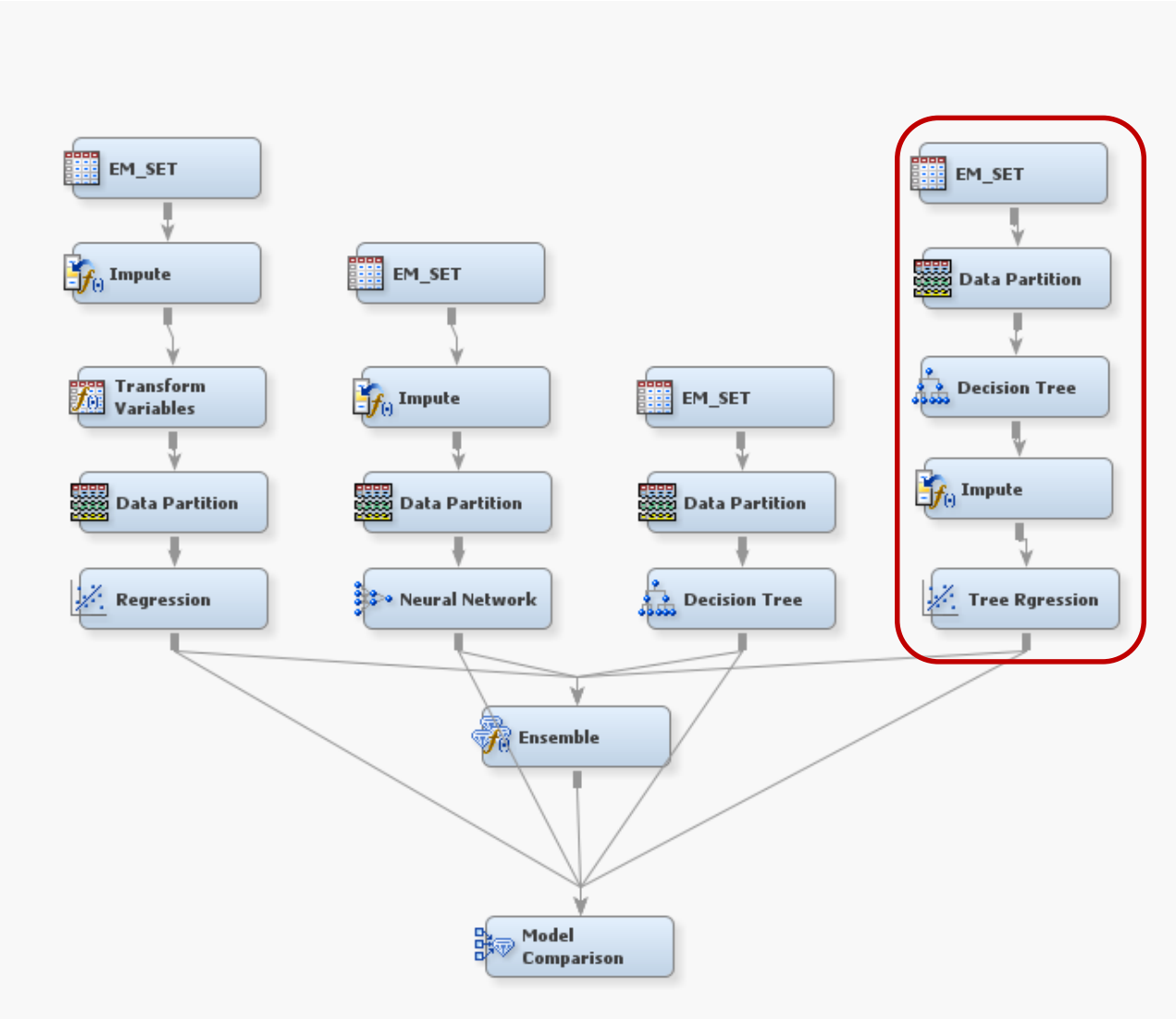


Figure 7. SAS® Enterprise Miner Diagram – Tree Regression

Since regression does not deal well with missing values, the impute node is used for performing missing value imputation. This node offers multiple imputation options for class variables and interval variables. One of the interesting methods of imputation offered is a tree imputation technique which is a versatile method of replacing missing values using an inbuilt decision tree. In essence, it tries to predict what the missing value is most likely to be based on all other available attributes.

As this project is a part of an internship, non-disclosure agreements keep me from divulging more interesting details of the various models that were developed.

TREATMENTS BASED ON PROBABILITY OF PAYMENT

Now that we have results from a fairly detailed statistical analysis, the next step is to take a business view of the results and take some decisions based on the same. Taking a step back to revisit the final aim of this analysis, we are trying to maximize our collections by way of increasing operational efficiency and deploying the right strategy for each account. So with this in mind, let us see how to use our new found knowledge for establishing treatments that are best suited for each account.

The below example takes in to account results from a few other models that were mentioned above in the paper and work in tandem with the PoP model we just went through.

Consider an account that gets a high probability of payment and an approximate month 4 of first payment. We can then use the below chart to assign a suitable plan of action for the account based on past knowledge. So if an account is highly likely to pay in month 4 from acquisition, we can confidently assign it to a call center through a non-legal channel which would keep our over-head expenses very low and return on investment high.

However, in case we have an account at hand that has a high probability of payment but the estimated month of first payment is well after 24 months, we will need to deploy a different treatment. In this case, we can subject the account to professional skip tracing and gather information about mortgage payments made in the recent past or a recent increase in credit rating. Using these indicators, a legal rout can be adopted based on the fact that such positive financial activity indicates that the customer is financially sound and can be profitably sued. However, the efficiency of skip tracing at this point becomes crucial considering the legal route is very expensive to go through.

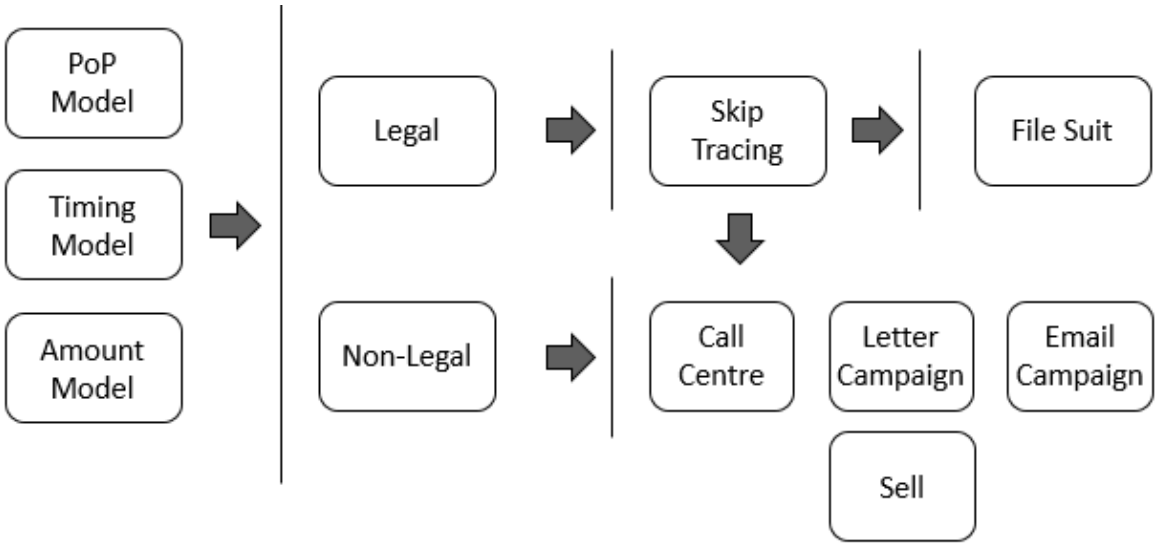


Figure 8. Treatments

CONCLUSION

With banks selling less delinquent accounts and at higher prices, the margins for debt collection companies are slim to work with. The debt collection industry is maturing in many aspects and it is vital for analysts to leverage ways that give the company an edge over others. Analytics is bound to be one of the most important tools that will aid better decision making. As described above, data is abundantly available and more can be used (for example customer data from mobile carriers) to support any analysis that can be conceived.

REFERENCES

- Collections Information: Facts and statistics about third-party debt collection.
  http://www.acainternational.org/products-collections-information-5431.aspx
- US Census
  http://www.census.gov/
- Decision Tree Validation: A Comprehensive Approach
  http://www2.sas.com/proceedings/sugi30/256-30.pdf

ACKNOWLEDGMENTS

I would like to thank my professor Dr. Goutam Chakraborty for being a great guide and pushing me to bring out my best. I also thank my colleagues at SquareTwo Financial – Eric Hayes, Annette Girmus Orford, Harold Dickerson and Thomas Waldschmidt for taking me under their wings and guiding me as their intern.

RECOMMENDED READING

- Enhancing delinquent debt collection using statistical models of debt historical information and account events
  US 7191150 B1
  http://www.google.com/patents/US7191150

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Karush Jaggi
Risk Analyst, AFS Accpetance
Ft. Lauderdale, FL – 33309
405-762-1946
karush.jaggi@okstate.edu

Thomas Waldschmidt
SquareTwo Financial
Denver, CO - 74075
720-217-6046
thomasw2101@icloud.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.