Predicting Fraud Cases for MoneyGram International

Michael Olheiser

July 28, 2018

1 Introduction to Problem

MoneyGram International is a financial services company that wires money across the world through their agents. These agents include physical locations such as banks, supermarkets, internet cafes, etc. Unfortunately, agents can be victims of social engineering that results in a loss of revenue. Social engineering fraud is the method used by fraudsters to convince an agent over the phone that they are from technical support to steal money from the agent. MoneyGram has over 300,000 locations which makes fraud training time consuming and expensive. This project is designed to target agents that may be the most vulnerable to fraud in the future so that fraud training efforts can be concentrated and reduce cost. The result should save MoneyGram resources and money devoted to fraud prevention.

2 Proposed Solution

The expected solution of this project is to build a predictive model using data from MoneyGram databases, the US Census, and FBI. The hypothesis is that information about the agents location such as crime rate, poverty rate, education, and population will be significant predictors of fraud. The scope of this project will only include agents in the United States that have the ability to send Money Transfer transactions. The data will be analyzed for significant variables and then used to train and test various predictive models. The resultant models will be analyzed for effectiveness in detecting fraud and reducing cost for MoneyGram.

3 Data Preparation

3.1 Gathering

The initial data set is pulled from MoneyGram databases that contains information about their agents such as location, credit, name, creation date, account blocks, software type, and more. Additional data was gathered from US Census and FBI websites that includes population, education, poverty, and crime per location (zip code, city, or county). This external data is joined to the MoneyGram data set by the agent's location. The resulting data set is 73,018 rows with 40 fields where each row represents a Point of Sale system belonging to a MoneyGram agent.

3.2 Cleaning

The raw data set contains many agents that are not applicable to the scope of this project therefore the data was filtered to only include agents with a specified agent type and the ability to complete Money Transfer transactions. Furthermore, the data grouped all Point of Sale systems together by agent since the scope of this project is to prevent fraud at agent locations and not individual MoneyGram workstations . The result is 8,313 agents located in the United States that will be used to train a test a predictive model.

The external data from the US Census and FBI also needed extensive cleaning. The data was joined on various location keys such as Zip Code, City, and County of the agents' location that were formatted to ensure

the join was successful. The data from FBI and the US Census did not have data for all agents so instead of excluding those agents from the model, the nearest location was used to fill in the missing values. If the nearest city was over 100 miles away or the nearest County was over 200 miles away, then the data remained missing. The resulting missing data is reviewed in section 4.3.

4 Data Analysis

The tidy data set contains approximately 8000 agents that will be analyzed for patterns that could potentially predict fraud. This section will cover location, significance tests of predictor variables, correlations, and a review of the missing values.

4.1 Plotting Location

This project is limited to Money Transfer agents in the United States. Figure 1 shows a geographic scatter plot of all agents considered in the project where red dots are agents that have been victims of fraud and black dots are not. The plot suggests that the incidence of fraud correlates with the location of other (non-fraud) agents so there is no particular place in the United States that fraudsters seem to target. The demographic information of these locations may be important and is explored in the next sections.

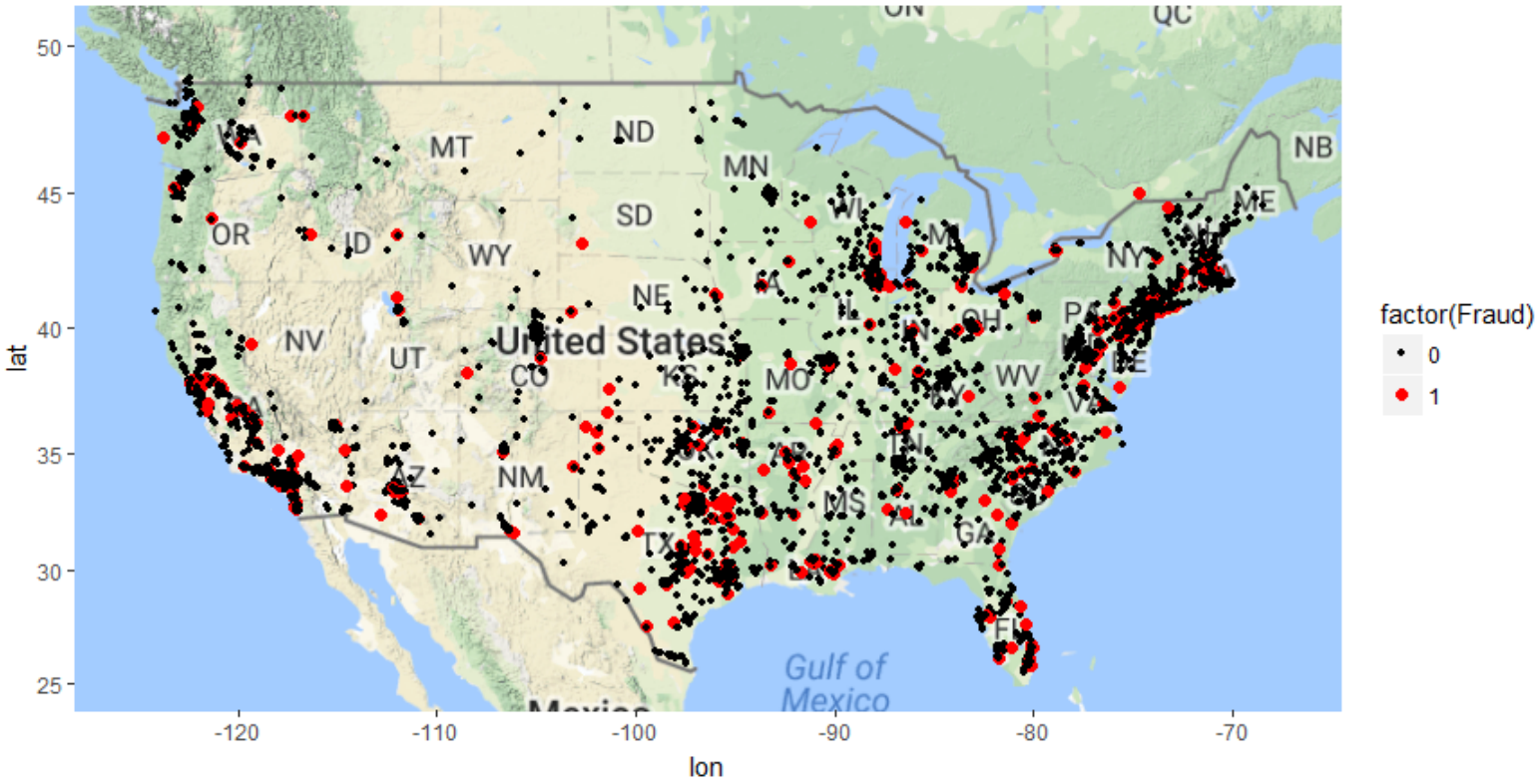


Figure 1: United States Fraud Map of MoneyGram Agents

4.2 T-test Results

A Two Tailed T-test was applied to each field to determine if there are significant differences between agents that have experienced fraud and ones that have not. The results show significant differences for County Population, Poverty Rate, Mean Credit, and Account Blocks. The latter two were masked due to being sensitive information that is not public outside of MoneyGram. These two fields, Mean Credit and Account blocks, had the smallest P value meaning the differences were highly significant suggesting that they may be effective predictors of fraud.

Table 1: T-test results

Field	μ_f-fraud	$\mu_{!f}$-!fraud	σ_f-fraud	$\sigma_{!f}$-!fraud	P-value
City Crime per capita	.0780	.0747	.0451	.0383	.151
County Pop	1,589,980	1,834,901	2,327,787	2,444,574	.0472
City Pop	426,871	512,263	953,248	1,037,160	.118
Zip Code Pop	24,319	51,574	21,911	20,339	.194
% Poverty	16.4	15.8	5.22	5.14	.0266
Median Age	35.4	35.3	5.37	5.80	.572
% HS Graduate	85.5	85.6	5.61	5.87	.838
% Bachelors	30.9	31.7	9.22	9.63	.0812
% Advanced Deg	11.6	12.0	4.72	4.89	.160
# Agents in Zip	29.9	30.5	24.5	23.3	.618
Mean Credit	*	*	*	*	<.001
Acct Blocks	*	*	*	*	<.001

4.3 Missing Values

This section will consider missing values to help determine which variables are best suited for a predictive model. The count of missing values are seen in Table 2. In this table, its obvious that City Population would be a difficult predictor for logistic regression therefore County or Zip Code Population can suffice. Section 4.4 will explore the correlations between the population fields and the rest of the data set.

Table 2: Missing values

Field	Missing Values
City Crime per Capita	128
County Pop	396
City Pop	1479
Zip Code Pop	20
% Poverty	164
Median Age of City Pop	20
% HS Graduate	151
% Bachelors	151
% Advanced Degree	151
# Agents in Zip	0
Mean Credit	0
Acct Blocks	0

4.4 Correlations

Table 1 shows 9 out of 13 fields with p-values less than 0.10 that may be predictors of fraud when building a logistic and CART model. Table 2 has shown that fields City Population, and County Population may be difficult to use in a Logistic regression due to the number of missing values. This section will look at a correlation table to further validate the use of each field in predictive modeling.

Table 3: Correlation values

	Crime	CntyPop	CityPop	ZipPop	Poverty	Age	HS	Bach	AdvDeg	Agents	Credit	Blocks
Crime	1.00	-0.13	0.01	-0.11	**0.30**	-0.04	-0.02	-0.20	-0.18	-0.02	0.06	-0.04
County Pop	-0.13	1.00	**0.40**	**0.36**	0.05	-0.17	**-0.40**	0.05	0.00	0.27	-0.01	0.01
City Pop	0.01	**0.40**	1.00	0.19	0.13	-0.14	-0.18	0.21	0.22	0.18	0.06	0.00
Zip Pop	-0.11	**0.36**	0.19	1.00	0.00	**-0.32**	-0.26	0.08	0.06	**0.61**	-0.01	0.03
Poverty	**0.30**	0.05	0.13	0.00	1.00	-0.15	**-0.54**	**-0.53**	**-0.44**	0.05	0.01	0.00
Median age	-0.04	-0.17	-0.14	**-0.32**	-0.15	1.00	0.26	0.02	0.04	**-0.34**	0.04	-0.04
HS Grad	-0.02	**-0.40**	-0.18	-0.26	**-0.54**	0.26	1.00	**0.54**	**0.49**	-0.22	0.01	-0.04
Bachelors	-0.20	0.05	0.21	0.08	**-0.53**	0.02	0.54	1.00	**0.95**	0.05	-0.01	0.01
Adv. Deg.	-0.18	0.00	0.22	0.06	**-0.44**	0.04	**0.49**	**0.95**	1.00	0.04	0.01	0.02
Agents	-0.02	0.27	0.18	**0.61**	0.05	**-0.34**	-0.22	0.05	0.04	1.00	-0.06	0.06
Credit	0.06	-0.01	0.06	-0.01	0.01	0.04	0.01	-0.01	-0.02	-0.06	1.00	-0.25
Blocks	-0.04	0.01	0.00	0.03	0.00	-0.04	-0.04	0.01	0.02	0.06	-0.25	1.00

Correlation values greater than or equal to 0.30 are bold and underlined in table 3. As expected, the same type of fields such as population, education, and credit contain correlations within them such as City Population and County Population. In these cases, modeling will avoid using predictors of the same type.

4.5 Data Analysis Conclusion

The exploratory data analysis revealed that many of the attributes in the data set were not significant to distinguish fraud agents from non-fraud agents. More specifically, these attributes were geographic location, city population, crime, and education rates. It was anticipated that fraud agents would be located in smaller, impoverished areas that are more prone to crime and low education rates but this is not necessarily true. The fields that had the most significant difference between fraud and non-fraud agents were taken directly from MoneyGram databases, that is Mean Credit and Account Blocks. Neither of these fields had any missing values making them suitable candidates in a predictive model. Although County Population and Poverty Rate were also significant, it is not clear if these will be predictors of fraud when used in conjunction with other attributes.

Fraudsters may be consciously selecting agents based on certain criteria that this report has not analyzed. Also, our data set does not contain the agents that fraudsters fail to socially engineer so its impossible to know if we are identifying the criteria that is consciously being used by fraudsters . Given the nature of social engineering attacks, we can reasonably assume that fraudsters are making multiple calls every day to attack agents thus it is possible that some of the attributes explored in this section could predict fraud. The next section will explore predictive modeling to determine which agents would be victims of fraud and what attributes contribute to that prediction.

4.6 Summary

5 Predictive Modeling

This section summarizes the outcomes of three different types of predictive models using the most significant predictor variables that were identified in section 4. The three types of models used are a Logistic Regression, CART (Regression Tree), and Random Forest. The predictors were kept the same for each model and were selected through experimentation and knowledge built in the exploratory data analysis.

5.1 Logistic Regression

In general, a regression measures the relationship between explanatory and outcome variables. In this report the outcome variable is logical (Fraud or Not Fraud) hence the use of a Logistic Regression versus a Linear Regression. The explanatory variables are Mean Credit, Account Blocks, Poverty Rate, and Median Age. Each explanatory variable has weight and significance level that determines how much influence the variable has in regards to distinguishing fraud from non-fraud. Below is the breakdown of each variable and coefficients.

Predictor	**Coef.**	**P-value**
Intercept	-4.685	<.001
Poverty	2.613e-2	.0111
Credit	-1.347e-5	<.001
Acct Blocks	9.086e-1	<.001
Median Age	3.634e-2	<.001

Given the relatively low P-values, there is likley a relationship between fraud and the predictors chosen. The credit value of the agent has the only negative coefficient amongst the 4 predictors meaning that a lower credit is related to more fraud. The other three predictors had a positive relationship with the incidence of fraud.

5.2 CART

The second model that was built was a CART (Classification and Regression Tree) model. In general, a decision tree uses predictor variables to decide which bucket an observation falls into. In this case, the observations are agents and the model is deciding the probability of fraud. An algorithm creates a downward flow chart that compares the observation against a series of logical questions until the final result shows the probability of fraud. The image portraying the decision tree is located in the appendix as figure 2.

The first decision node splits on whether or not the agent had at least one account block and then splits again multiple times based on the credit value. This indicates that these two attributes are the strongest predictors which is in line with the t-test results from the previous section. Median Age and Poverty Rate are slightly less important for predicting fraud but still provide a more granular prediction in each split. The splits of this tree suggest that fraud is associated with a lower credit, higher median age, and higher poverty rate.

5.3 Random Forest

A Random Forest is an ensemble of CART models that averages the results. After reviewing the decision tree (see: Appendix) produced by the CART model, its possible that there may have been over-fitting due the number of splits with very close values such as Median Age ¿= 41 splitting into Median Age ¡ 40 and Median age ¡ 42. The model is deciding there is a difference in fraud whether the median age is 40, 41, or 42 which to common knowledge probably would not make any difference. A Random Forest creates multiple models to try and correct for this over-fitting.

5.4 Metrics

The following table shows 3 different models with varying probability thresholds that were selected using an ROC curve. The last "model" is simply guessing where it is assumed that each agent has a 50% chance of fraud occurring to create a baseline for model comparison. The largest or smallest value for each metric is in bold. The four predictor variables remained the same for all models.

5.5 Model Selection

Selecting the best model is not a cut and dry process as it depends on the situation that the model will be applied to. The obvious goal in this case is to reduce the amount of fraud however the ultimate goal is to reduce costs

Table 4: Predictive Model Metrics

Model Type	Threshold	TP	TN	FP	FN	AUC	Sens.	Spec.	Acc.	FP Rate	FN Rate
Log. Reg. 1	.03	**102**	595	1338	**9**	.739	**.919**	.308	.341	.692	**.081**
Log. Reg. 2	.04	87	914	1019	24	.739	.784	.473	.490	.527	.216
Log. Reg. 3	.05	77	1132	801	34	.739	.694	.586	.591	.414	.306
Reg. Tree 1	.02	**102**	604	1329	**9**	**.810**	**.919**	.312	.345	.688	**.081**
Reg. Tree 2	.03	79	1082	851	32	**.810**	.712	.560	.568	.440	.288
Reg. Tree 3	.04	69	**1227**	**706**	42	**.810**	.622	**.635**	**.634**	**.365**	.378
Rand. For. 1	.02	101	622	1311	10	.733	.910	.322	.354	.678	.090
Rand. For. 2	.03	82	1040	893	29	.733	.739	.538	.549	.462	.261
Rand. For. 3	.04	77	1151	782	34	.733	.694	.595	.601	.405	.306
Guessing	.50	56	967	966	55	.500	.505	.500	.500	.500	.495

for MoneyGram so the selected model will need to find a balance between False Positives and True Positives. For every False Positive, MoneyGram would be providing additional training to agents that don't actually need it thus creating more cost. For every true positive, MoneyGram would be reducing fraud costs if we assume that the training is effective. With that in mind, we should select a model where the Sensitivity and Specificity are closest while leaning slightly to a higher sensitivity. The models that fit this need are the Logistic Regression 3, Regression Tree 2, and Random Forest 3. The next section will break down a hypothetical situation to assess the financial impact of the models.

6 The Data Story

Through analysis and experimentation, the following fields were chosen to build predictive models: Credit, Median Age, Account blocks, Poverty Rate. Each field was selected based on a number of factors covered in Section 4 including significance from the T test, missing values, improvement of the model, and the story each attribute tells. The model was intended to make conceptual sense in regards to social engineering fraud.

Fraudsters that use social engineering as an attack vector are successful because of two reasons: probability and the human brain. Fraudsters will call many people every day trying to manipulate them for personal gain knowing that they will fail most of the time. However, fraudsters are consistent and diligent and eventually they find a person that is more likely to be compromised by social engineering. The reasons why certain people fall for these scams is not so clear. MoneyGram agents are no stranger to this attack.

This project was designed to determine why certain agents fall victim to fraud over others. Fraudsters typically pretend to be from technical support which means they have some story about how the agent has a technical problem that needs resolution. The number of times an agent has had their account blocked is a possible metric to identify agents that have trouble with MoneyGram systems and are more likely to have a technical problem when a fraudsters calls. It is apparent in the Regression Tree model that the most significant decision for predicting fraud is whether or not the agent has had their account blocked in the past. The difference between fraud and non fraud agents was at least one account block. This alone does not tell the whole story however. Agents that have fallen victim for these attacks tend to have much lower credits as seen in the regression tree. On average, agents that have not had fraud have over 200% higher credit than Agents with fraud. A lower credit means that they do not transact often and likely to be less familiar with MoneyGram systems thus more susceptible to social engineering.

The best defense of social engineering is education. However, there a multiple factors that influence the education. This project anticipated that some demographics may determine whether or not an agent has education available to protect their employees from fraudsters. Poverty rate was a significant predictor for fraud in each model. Agents that are located in impoverished areas are more likely to not have the financial resources to keep employees nor provide consistent education to protect their employees. The average Poverty rate for fraud agents was 5% higher than agents without fraud; a small but significant difference (P-value of 0.026). Furthermore, it is not unknown that older people are not as familiar with technology and are more susceptible to social

engineering scams. The results of the predictive models show that a higher median age of the population is a significant predictor of fraud. The t-test results showed less than a year difference on average (35.4 vs 35.3) but was significant to predict fraud when used in a model with the other fields. This highlights that there is not necessarily one obvious attribute that distinguishes fraud consistently. It is known that fraudsters change their tactics and the lack of significance among the analyzed data may be supporting that fraudsters are dynamic and able to attack many different types of agents.

7 Operational Cases

The previous section explored model selection by using common metrics to assess the effectiveness of those models. In a business setting, the ultimate metric is profit. Here we will assess the amount of money this model can save MoneyGram with some pre determined costs. It is important to note that the values of cost used in this section are hypothetical and in no way represent MoneyGram's actual operations.

Table 5: Example Data for Financial Impact Assessment

Total Agents in Scope	2044
Fraud Loss per Case (*Fraud*)	$3,000
Training Cost per Agent (*Training*)	$200

7.1 Case 1: Operations Remain As-Is

First we will assess the situation without using a model and training. In this case, MoneyGram will continue fraud operations as usual and will not provide additional fraud training to agents. Based on the testing data set, there are 111 fraud cases. We will assume that an average fraud case is $3000 thus $330,000 would be the total cost by continuing the current fraud prevention operations.

$$\text{Total Cost} = \$330{,}00$$

7.2 Case 2: Fraud Training for All Agents

In this second case, MoneyGram would provide fraud training to all agents in this data set which is 2044. In this case we assume the training is 100% effective and no fraud occurs at these agents therefore the total cost of this case is attributed to training. We will assume the cost of training is $200 per agent, so the total cost for all agents is $408,800 which is $78,800 more than the cost of fraud itself therefore this additional training program would not make much sense to complete.

$$\text{Total Cost} = \$408{,}800$$

7.3 Case 3: Training for Randomly Selected Agents

In this case, MoneyGram will reduce the number of agents to train by randomly guessing which agents would be victims of fraud with a 50% chance of fraud occurring. MoneyGram would provide training for 1022 agents that costs $204,400 total while missing 55 fraud cases that would cost $165,000 which totals $369,400. Similar to case 3, the costs of this approach are higher than continuing operations as usual therefore this would not be a cost effective option for MoneyGram.

$$\text{Total Cost} = \$369{,}400$$

7.4 Case 4: Training for Agents Selected By Model

In the last case, MoneyGram would use a predictive model to select which agents should receive the fraud training. In each model the agents that need to receive training are the sum of the true positives and false positives. There is also the cost of fraud that is sum of false negatives multiplied by the cost of each fraud case. The expression of total cost is shown below. The most effective model for MoneyGram will be the one that can reduce the total cost the most.

$$TotalCost = (TP + FP) * Training + FN * Fraud \tag{1}$$

8 Financial Impact of Model

To measure the financial impact of each model we will need to select some hypothetical values to calculate the total operational cost of deploying each model. Table 5 summarizes a possible scenario of costs for a fraud case and training cost for each agent. In this case, the testing data set of 2044 agents will be the example set used in this analysis.

The base line cost will be $330,00 that was calculated from case 2 that is, operations remain as-is since this was the cheapest option without using predictive modeling. Using equation (1), we can calculate the total cost of each model and subtract that from our base line cost of $330,000 to determine the savings that each model provides. Table 6 below summarizes these costs for each model.

Table 6: Financial Result of using Model

Model	Threshold	Agents to Conact	Total Cost	Savings
Log. Reg. 1	.03	1440	$315,000	$18,000
Log. Reg. 2	.04	1106	$293,200	$39,800
Log. Reg. 3	.05	878	$277,600	$55,400
Reg. Tree 1	.02	1431	$313,200	$19,800
Reg. Tree 2	.03	930	$282,000	$51,000
Reg. Tree 3	.04	775	$281,000	$52,000
Rand. For. 1	.02	1412	$312,400	$20,600
Rand. For. 2	.03	975	$282,000	$51,000
Rand. For. 3	.04	**859**	**$273,800**	**$59,200**
Guessing	.50	1022	$369,400	$(36,400)

Random Forest 3 is the model with the most cost savings which is one of the models that was selected as a suitable candidate in section 2.2. In fact, the three models selected in section 2.2 were in the top four of highest savings. The Regression Tree 3 was not selected in section 2.2, but still had the second highest cost savings mostly due to having the lowest false positive rate. Since the costs are hypothetical, its important to note that each model may be ranked differently depending on the actual operating costs. Keeping this is mind, it would still be advised to select a more sensitive model in order to cover the most fraud agents.

9 Conclusion

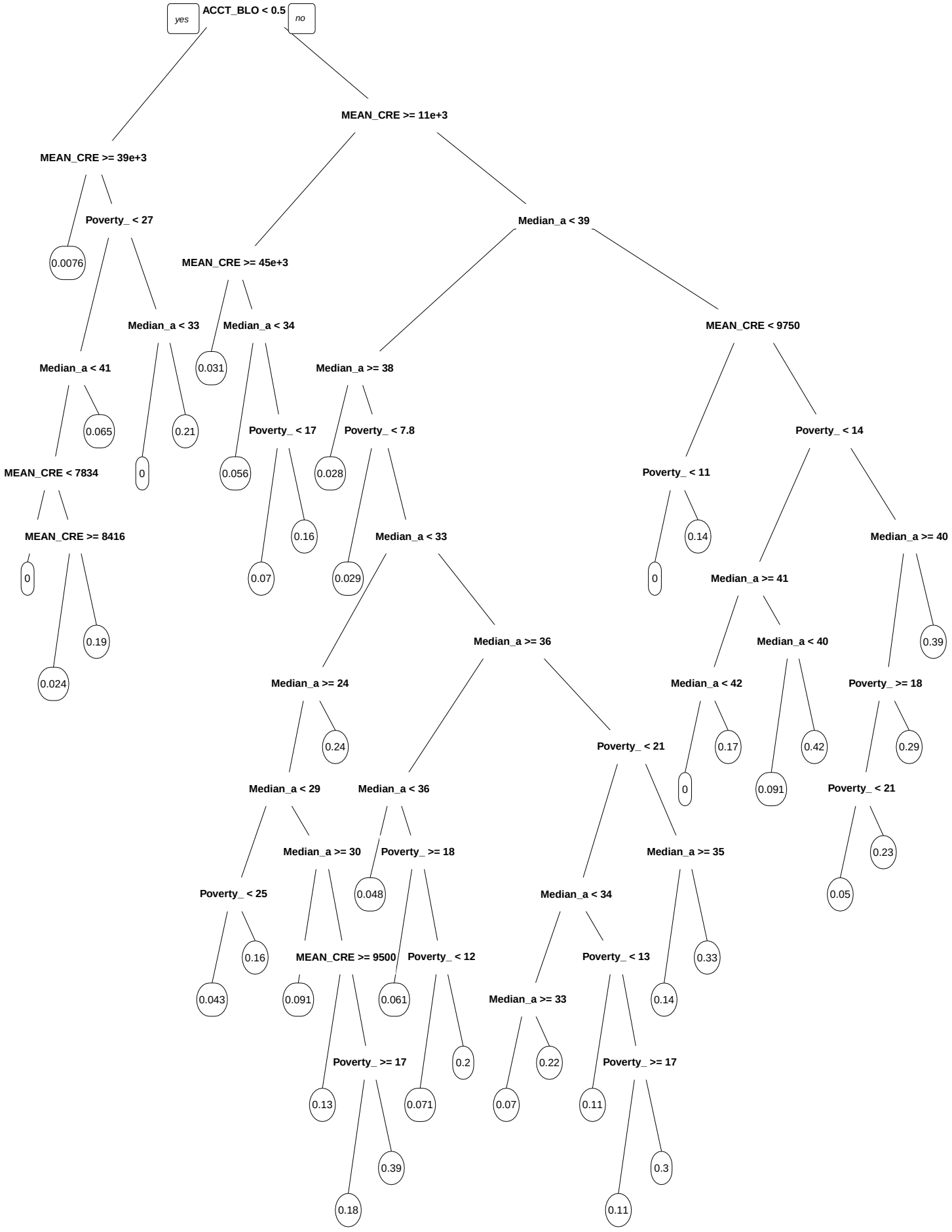
Three unique models were tested on a data set of 2044 MoneyGram agents using 3 different probability thresholds each. The ultimate goal for MoneyGram is to reduce costs therefore the best model will likely have a small difference between sensitivity and specificity. Possible candidates includes Logistic Regression 3, Regression Tree 2, and Random Forest 3. The financial impact analysis showed all three models in the top four of costs savings along with Regression Tree 2 that was not initially chosen due to having a higher specificity. It is still advisable to choose a slightly more sensitive model in order to cover the most fraud cases in the future.

One cost that is not directly measured by dollar amounts, is brand reputation and confidence. Reducing fraud

might create a stronger customer base for MoneyGram so reducing operational costs would not be the sole factor in determining the model. With this in mind, Random Forest 3 is the most suitable model to deploy. This model had the highest reduction in costs while still predicting 77 out of 111 fraud cases correctly. Should MoneyGram decide that protecting the company from fraud attacks weighs more heavily than the operational costs, a more sensitive model can be chosen and still provide a cost reduction as no model had a negative cost reduction.

This report examined three different types of models using the same four predictor variables. Further research should be done with Random Forests or more sophisticated models using the other variables. It might be possible to build a model without being affected by correlations and missing values as much as a logistic regression.

10 Appendix


yes
ACCT_BLO < 0.5
no
MEAN_CRE >= 39e+3
MEAN_CRE >= 11e+3
Poverty_ < 27
Median_a < 39
0.0076
MEAN_CRE >= 45e+3
Median_a < 33
Median_a < 34
MEAN_CRE < 9750
Median_a < 41
0.031
Median_a >= 38
0.065
0.21
Poverty_ < 17
Poverty_ < 7.8
Poverty_ < 14
MEAN_CRE < 7834
0
0.056
0.028
Poverty_ < 11
0.16
Median_a < 33
0.14
MEAN_CRE >= 8416
0.07
0.029
Median_a >= 40
0
Median_a >= 41
0
0.19
Median_a >= 36
0.39
Median_a >= 24
Median_a < 42
Median_a < 40
0.024
0.24
Poverty_ < 21
0.17
0.42
Poverty_ >= 18
Median_a < 29
Median_a < 36
0
0.091
0.29
Median_a >= 30
Poverty_ >= 18
Median_a >= 35
Poverty_ < 21
0.23
Poverty_ < 25
0.048
Median_a < 34
0.05
0.16
MEAN_CRE >= 9500
Poverty_ < 12
Poverty_ < 13
0.33
0.043
0.091
0.061
Median_a >= 33
0.14
Poverty_ >= 17
0.2
0.22
Poverty_ >= 17
0.13
0.071
0.07
0.11
0.39
0.3
0.18
0.11