Marc Burt
marc.s.burt@marcsburt.com
254-423-1457

Saylent Data Analysis:

ETL:

Assumptions: This data came from a relational database. The data is a sample set given from the bank, not a full set. I will go through the ETL process with a goal of transforming the data to use for predicting fraudulent charges, but will discuss other compelling uses for this data in later sections. I also need to note that ETL decisions should be made by committee, since I would not be the only one accessing this information going forward.

Credit Card Numbers: There are currently only 13 unique credit cards listed. I assume in real data there would be a similar amount of credit card numbers than unique names. Specific sections of the credit card numbers could provide attributes of the card i.e type of card, bank origin, etc.

Name: There are 180 unique names with all of the last names starting with an “A”. This is more interesting than useful. More useful would be to split the names into first, middle, and last name columns. I see a few names within the data that would have to be dealt with. Some examples are “M\xe4dchen Amick”, “Brooke Adams (actress)”, “Veronica Porch\xe9 Ali” and “Aaliyah” stood out to me as names that were misentered or misprocessed somehow. While the purpose of this report isn’t to point out the data’s validity, it’s not uncommon to have fields that need to be scrubbed more than others. I believe the name field is one of those.

SSN: There are 12 unique Social Security Numbers present. If the data was real, we would have more unique SSN’s then unique names. I would create a unique key of transaction data from some combination of last few numbers of a person’s SSN and some of the later digits of the credit card numbers. SSN by itself is an identifier which would not provide useful information in terms of predictive analysis.

Customer Address: The addresses seem to be PO boxes and physical addresses, both without city or state information. I would assume mailing and physical address information is differentiated in the bank’s database. If this were correct, the address would probably be less useful then it’s physical proximity to where charges took place. For example if I wanted to use addresses as a predictor for fraudulent charges, I would ask whether distance from home were a factor. Google has an API that serves JSON or XML that can take two addresses and gives back a numeric distance.

Transaction Method: This column has as many unique identifiers as there are rows of data. I am not familiar with this column and I would need to do more digging. There are values like “Pinned Purchase” that I understand at face value, but I am unfamiliar with what BIN #####

Marc Burt
marc.s.burt@marcsburt.com
254-423-1457

means. I believe that transaction type could have some bearing on fraudulent charges, but I would have to find some kind of categorical proxy for bin numbers.

Merchant: There are 12 unique merchants in the data. A few things that stood out are that 100% of the 187 transactions originating at BJ's and the 62 transactions from Flowers.com are fraudulent. I believe merchant data would be a great predictor of fraudulent transactions. I might look into categorizing these into types of merchants to minimize the dimensionality if the dataset were more robust. In addition I would want to the addresses of the locations that transactions happened so I can gather similar proximity data mentioned in the Customer Address section above.

Transaction Date and Settlement Date: This data is from 1-Nov to 19-Nov with an unspecified year. The range of fraud happening on a specific day is between 23% and 30%. Similar to the proximity within various addresses, I would want to split dates into days of the week, time between transaction and settlement date, and some kind of categorical method of differentiating between different times of the month.

Amount Completed: The range of data is from $5.00 to $336.00 with a mean of $128.00 and a median of $75.00. This is a field that does not need to be changed or transformed.

Balance Available: The range of data is from $600.00 to $40320.00 with a mean of $22584.00 and a median of $38040.00. This is a field that does not need to be changed or transformed unless I were using an algorithm that had difficulty with mixing continuous data with categorical. In this case I would bin these values into proportionate sizes.

Terminal Location: Only the first 56 rows of data have terminal location listed. This is some kind error in the dataset that would need to be cleared up. If this is zip code data, it could be very valuable in our predictive model.

Institution Name and Acquiring Institution Name: There are 30 unique categories of institution names. These variables would need to be factored in a way a machine learning algorithm would accept. 30 is at the upper limit in terms of dimensionality, but still within what I would deem as acceptable.

Is Fraud: Out of all the transactions in the data set, 27% are fraudulent. This is highly unlikely, and may be evidence of some kind of sampling error. I just attribute the unusually high number to this data being fake which I take at face value.

Marc Burt
marc.s.burt@marcsburt.com
254-423-1457

Insights From Data:

Perhaps the most glaring insight in this data would be what factors could be used to predict fraudulent charges. However, there are other insights we could gain from this data as well. I would be interested in knowing how far away most of my customers travel to use their cards. Are there certain stores that people shop at in succession more than others. What are the big predictors on whether a customer uses their card at an ATM or with a merchant. Does location have a big impact on bill size?

Steps Taken: Predicting Fraud

Technology Stack: During these steps I will do most of my preprocessing in R or SQL. SQL lends itself more toward organizing data, but R lends itself to factoring categorical data, splitting training/testing sets, and quickly rendering plots with ggplots. I might even prototype some models in R because I personally enjoy the syntax and ease of use in a lot of the machine learning packages offered. However, given that models are being put into production, I would probably use Python for most of the later steps. It's just easier to transition Python to production than R. I might even look into distributed file systems like Spark on a Hadoop cluster for production purposes, but that is outside the scope of this report.

Steps:

- Assuming that most of the preprocessing steps were taken care of, I would start with sampling data. I would assume that the full dataset is hundreds of thousands of rows long, perhaps more. I picture either being able to take a random sample of the full amount of information, or giving the bank instruction on how to sample the data. This really depends of level of access the bank is willing to give. Right now the data seems to be a subset of a much larger dataset from people whose last name start with an "A", which is clearly not random.

- Next I would start to prototype using models like Decision Trees or Naive Bayes with k-fold cross validation to minimize sampling error within training and testing sets. Decision Trees are very easily interpreted and are easily prototyped. Naive Bayes would lend itself as a second model that is easily interpreted, but probably less accurate unless we discretize some if the continuous variables. These effects will become more prominent as we add more features. In both instances I would have to deal with high dimensional categorical data. The more dimensionality, the more I expose models to overfitting. Either I find proxies for categorical data, or remove them from the dataset. I

outlined some categories in the ETL section that would need some dimensionality reduction. R only allows 32 factors within most of the modeling packages.

- At this stage I would start adding features to the models. Features like proximity data that I mentioned in some of the ETL sections. I would leverage other financial data from individuals that are probably hidden in another table that could be linked to. I could split up the month and days into days of the week which helps with dimensionality reduction. As I add features I test them within my prototyping models. One thing I would need to keep in mind is adding features that are correlated with existing features. I could test both individually to see which one has more predictive qualities and toss the other.

- Next I look into more sophisticated models that might trade understandability with sophistication. Sometimes the extra level of sophistication does not make a better predictive model, but it's always worth testing. These models include Random Forest, Neural Nets, or Support Vector Machines with different kernels. Again, these models are not as easily interpreted, however they could be more valuable.

- Measurement of accuracy can be done a few different ways depending on the business need. Area under the receiver operator curves (AUROC) serves as a good baseline measurement between different models. However, given different purposes, I would use other measurements. For example, if I were sending my credit card user a text message if their card is possibly being used in a fraudulent charge, I am more concerned with reducing false negatives than false positives. What is the harm in sending a text as a warning even if the charge isn't fraudulent? However, if my bank shut off those credit cards that my predictive model output as being used fraudulently, then I would need to be sure that I was correct. In this case, I would place much more emphasis on reducing the amount of false positives.

- Of course, given that these models would be used in production, other factors such as resource usage and efficiency would be taken into account. This is where I would work closely with the development team to gain a better understanding of where this model would fit into their production schema.

Other Holes in The Data:

I believe that adding features from other areas of the relational database would lend more insight into what holes exist. Just a few that would be useful in this analysis and possibly future models would be:

- Age: Low dimensionality, easily tested and applied
- More Transaction Data: Not only how much, but how many items were purchased
- Marital Status: Low dimensionality

Marc Burt
marc.s.burt@marcsburt.com
254-423-1457

- Home Address: For more proximity measures
- More specific information on merchants

The list could go on, but often as a data analyst/scientist, you work within the constraints of the data you have access to. This isn't limited to data gathered internally, but also externally through various various other sources.