



Building A Fraud Detection Model

By Dr. Arthur L. Dryver

Email: Dryver@gmail.com

URL: <http://as.nida.ac.th/~dryver>

National Institute of Development
Administration (NIDA)

August 11th 2005



First, Who Am I?

■ Work Experience:

- I am a lecturer at NIDA within the School of Applied Statistics and I also teach in the School of Business Administration. I started working at NIDA in October 2003.
- I worked in the US for 4 yrs within consulting. Companies I worked for:
 - **Experian, AnaBus, and PricewaterhouseCoopers (PwC)**
 - While working as a consultant I performed analyses on data from various industries.
 - **For several projects I had to work with multiple files with over a million records.**
- At Experian I had to create fraud detection models for clients.

■ Education

- The Pennsylvania State University
 - Ph.D. in Statistics 1999
 - Dissertation Topic: Adaptive Sampling Strategies
- Rice University
 - BA in Mathematical Sciences/Statistics 1993



Data Mining

General statements about data mining.



Overview Of Data Mining

- Unfortunately, there is not only one way to look at data mining, and the differences among the data mining literature highlight this fact.
 - I will give you my overview for a typical project.

The Data

The Technique

The Presentation



The Data

1. What data do you have and can the data answer the questions you have?
 1. Do you have the data necessary to answer your questions.
 2. Note: The data that you have to use partially determines the technique you will use for data analysis.
2. Garbage In Garbage Out (G.I.G.O.)
 1. This is very important. What it means: You cannot expect to get good, reliable results with bad data.
 2. In other words: If the data is not good, not accurate, not reliable, etc. you cannot trust the results.



The Technique

- There are many data mining techniques.
 1. There is often more than one technique that can be used to answer the same question.
 1. The results from the different techniques often do not differ as much as one might believe.
 2. The technique is partially determined by the data you have.
 2. Many techniques within the data mining literature can also be found within standard statistics text books.
 - Data mining and statistics are both used to analyze data in order to gain useful information.



The Presentation

- The presentation is a very important part of data mining. Sometimes it can be the most important part of data mining.
 1. A good presentation should support the findings not just mention the findings.
 1. The supporting statistics, and graphs within the presentation can help people understand or confuse people.
 2. Management will often rely on the presentation to understand the findings from data mining.
 1. Management needs to trust the findings, if the findings are presented poorly, it is difficult to trust the findings.
 2. A poor presentation can even cause projects to fail. Management will not implement what they do not trust nor understand.
 2. Unfortunately, many statisticians and computer scientists are lacking in this critical area.
 - They tend to merely look at the results and the numbers in the computer output.
 - This makes many data analysis projects not as successful as they should be.
 - The poor presentation, explanation often leaves management unclear on how to understand and proceed with the findings from the project.



Building A Fraud Detection Model

One project that uses data mining.



What is Fraud?

- Fraud:
 - “A deception deliberately practiced in order to secure unfair or unlawful gain.” (Dictionary.com)
- There are different types of fraud committed within the credit industry.
- The type of fraud we want to model determines the data needed.
 - This presentation will focus on the typical application credit card fraud.
 - This will be defined as someone pretending to be someone else in order to obtain credit and never pays.
 - We **will not cover** bust out fraud.
 - Bust out fraud is when an individual fraudulently obtains credit. The individual is a good customer for a period of time to obtain higher levels of credit. When the credit limit is high the individual will borrow the maximum amount possible and then not repay.
 - We **will not cover** other types of fraud as well.
 - Should you have questions on any type of fraud please feel free to ask after the presentation.



How Do We Identify Fraud?

- How do we identify fraud?
 - How do we know the nonpayment is a result of fraud and not a result of a bad risk?
 - Credit Risk
 - “The risk that a borrower will be unable to make payment of interest or principal in a timely manner.” (Dictionary.com)
 - Sometimes an individual will obtain credit legally but be unable to pay, even the first payment.
 - This is known as first payment default (FPD) when an individual does not make the first payment, but did not commit fraud.
- How do we distinguish between fraud and first payment default?
 - Honestly, it is often difficult to distinguish between fraud and first payment default.



Fraud Vs. Risk

- Fraud and risk are trying to answer two very different questions.
 - As a result the data required to answer these questions differs as well.
 - For fraud we desire identification data and credit information.
 - Even the date of the data desired is different.
 - The most recent data is desired for fraud.
 - For risk we desire prior credit information.
 - Even the date of the data desired is different.
 - The flag (Good/Bad) should be present date, but the independent variables should be dated 6 to 12 months prior.



Creating A Good Statistical Model

- Many techniques used to create a good statistical model are useful to create a fraud detection model.
- A good statistical model starts with good data.
 1. Again the old saying garbage in garbage out (G.I.G.O.)
 - What is my point with this statement:
 - The message is that if the data collected is not good that the model is not expected to be good.
 - Why is this of concern in fraud detection models?
 1. We need to distinguish between fraud and first payment default. If we combine fraud with FPD it is like trying to create a single model to determine risk and fraud.
 - This can be done but it will create a model that does not work well on either risk or fraud. It is better to create two separate models instead.
 2. We need our independent data to also be accurate.
 - If our database determines the application contains incorrect information, it is important that this is not an error within the database.



Why Create Separate Models

- Why not create a single model to eliminate all potential bad clients?
 - Why separate FPD from fraud in the data?
 - Why not a single model to identify fraud and risk together?
- Benefits of Creating Two Models
 1. Clearly identify which factors are indicators of fraud and which are representative of risk.
 1. A large part of good model building is understanding why a variable belongs in the model. This is very difficult to do when you build a model to answer multiple issues, such as fraud and risk together.
 1. Only with a lot of model building experience on fraud and risk separately would you be confident in statements about the variables.
 2. Some variables may be driven by both fraud and risk, but determining which, fraud or risk, had a stronger influence on the variable selection would be difficult.
 2. The losses due to fraud and risk are not the same.
 3. The number of observations of fraud is often much lower than that of risk. Typical models built on risk use 10,000 observations of “bads” due to risk. Often it is difficult to obtain even 1,000 “bads” due to fraud.
 1. Creating a model using 500 “bads” from fraud and 9,500 “bads” from risk would create a model that focuses on risk.
 2. Creating a model using 500 “bads” from fraud and 500 “bads” from risk just to keep the numbers from fraud and risk equal is also not a good solution.



Separating Fraud and FPD

- As stated earlier it is important to differentiate between FPD and fraud when building a fraud detection model.
- In addition, we would remove the FPD data from the model building process.
 - In truth some FPD are probably unidentified fraud.
 - If we treat FPD as non-fraud when building the model we would have some frauds listed as both fraud and as non-fraud.
 - This is very bad to have when building a model.
 - The statistical model will contain variables to differentiate between frauds and non-frauds.
 - This will be more difficult to create if many frauds are labeled as non-frauds.



Data Required

- Some of the data needed for fraud detection is different from that of risk.
 - Important data on fraud detection tends to be identification information.
 - In the application for credit identification information is collected.
 - The identification information is then comparable to a large database containing people's identification information.
 - Difference between the identification information from the application and that of the database are signs of potential fraud.
 - When building a risk model, identification information is not needed. With a risk model, it is believed the person is who he say he is, but the concern is that he will not repay money borrowed.



Know The Data You Have

- Know your data.
 - What is the percent of success in the dependent (fraud) variable?
 - What are the values of the independent data?
 - Min, Max, Mean, Median.
 - **Default values**, are there any? How will you handle these values?
 - Example of a default value: Age unknown given a value of zero in the data.
 - **Outliers** – do they exist? How will you handle these values?
 - The handling of outliers will depend on how you will use your model. In practice often “capping” is used. Example, any number greater than the 99th percentile is set equal to the 99th percentile.
 - As with normal linear regression it is risky to extrapolate to values of the independent variable that weren't used in the model development.



What Technique To Use To Detect Fraud?

- Honestly as stated earlier there is more than one way to build a model to detect fraud.
 - The standard at my old company was logistic regression.
 - Logistic regression is used when there is a binary response variable, such as fraud.
 - Binary response means there are two possible outcomes. In our case fraud and not fraud.
 - Other possible techniques include decision trees and neural networks.
 - We felt and from some investigation that there was not an advantage to the other techniques.
 - In business: “Time is money”. In data mining we often do not have enough time to try all three and compare results.
 - On a few projects though we compared results of other techniques to logistic regression. There was no evidence that the other techniques were better than logistic regression.



Creating a Logistic Regression Model

- Create an Estimation and Validation Sample.
 - This step is very important when creating a model to be used in the future.
 - Validity – “The extent to which a measurement is measuring what was intended.” – Dictionary of Statistics B.S. Everitt. In other words does the model truly differentiate between success and failure?
- What is an Estimation and Validation Sample?
 - How many people have heard of a validation sample?
 - Oddly enough it was not covered much in graduate school, more briefly mentioned.
 - A validation sample is necessary when building a model to be used in practice.
 - Validations are discussed much more greatly in fields that apply statistics.



Creating a Logistic Regression Model

- What is an Estimation and Validation Sample - continued?
 - Estimation sample - The sample used to determine the parameters in the model.
 - Validation sample – The sample used to determine if the model produces consistent results. Will it perform well in practice, on another set of data? The validation sample is another set of data used to answer investigate this question.
 - Note: If the data is biased in general then the validation sample will not help in determining this. Example:
 - No women in all your data. It is not possible to know what will happen when the model is applied to all people. The validation sample has the same limitation as the estimation sample, thus the validation sample is not informative here.



Creating a Logistic Regression Model

- Now that we know what an estimation and validation sample is how do we create them?
 - The easiest sampling method is simple random sampling.
 - A more commonly used sampling design is stratified sampling.
 - Stratify your population into 2 groups, successes and failures.
 - When ample data is available sample 10,000 successes and 10,000 failures for the estimation and another 10,000 successes and 10,000 failures for the validation. Keep track of proportion of successes in your population relative to the number of successes sampled. Do the same for failures.
 - Often there is not enough data, often you will not have 10,000 frauds. Usually one group will have a lot and another group will have very few. For the group with a lot you need not sample beyond 10,000. This is an opinion, and it depends in part on how many variables you plan to use in your model.
 - In my opinion when data is rare, small in quantity:
 - I create a larger estimation sample than validation sample. Personal preference, haven't read anything but it is preferred in practice.



Creating a Logistic Regression Model

- Many people feel with modern computers sampling is not needed. Sampling is still needed:
 1. Without sampling you would only have an estimation sample and no validation sample.
 2. When dealing with millions of records, sampling can greatly aid in the speed of the analysis.
 1. Note: Many companies do not even have the latest and greatest in terms of computers. Many consultants work on their laptops, etc.
 3. Ultimately, when the model finished it is run on the entire dataset.



Variable Selection

- Variable Selection:
 - Often in practice we use brute force to create a model.
 - Example: We have 500 variables or more and then try all in the model.
 - We use Stepwise Logistic Regression to eliminate most of the variables and determine the best 15-20 or so most important variables.
 - Stepwise logistic regression is standard in most common software packages.
 - Honestly, in SAS for the speed we use a procedure called StepDisc first to come up with the first top 60 variables and then do Stepwise Logistic Regression.
 - Investigate the variables selected do they make sense.
 - For example: A mismatch with zip code on the application and zip code in the database should have a positive relationship with the presence of fraud. A negative relationship would make us question the results.



How Good Is the Model

- How do we know if the model is a good predictive model, or does it need more work?
 - First what is good?
 - Does the model distinguish/separate between the two groups (logistic 2 categories)
 - How do we tell if it is good?
 - Does the model validate well?
 - Do the statistics to test the model appear similar on the estimation and validation samples
 - Most important of all: Does the model distinguish/separate between the two groups (logistic 2 categories)
 - We will see what happens if we remove or change some of the less important variables in the model and compare results.
 - We will cover ways to determine how good the model created is in the following slides.



How Good Is The Model:

The KS Statistic

- What is the KS statistic?
 - It is the Kolmogorov-Smirnov two sample method
 - “A distribution free method that tests for any difference between population probability distributions. The test is based on the maximum absolute difference between the cumulative distribution functions of the samples from each population” – Dictionary of Statistics B.S. Everitt
 - A common statistic used to understand the predictive power of a model.
- How does it work?
 - Two cumulative distribution functions can be created, one from the successes and one from the failures. From logistic regression we estimate the probability of success and the probability of failure. Consider the probabilities of failure as the “random variable” then from this we can create two cumulative distribution functions one for the successes and one for the failures.
 - This will be illustrated on the next slide.



The KS Statistic

1st what is the KS statistic?

2nd does it look like this logistic model is predictive?

Take a minute or two to think about this question.

A 10 is the lowest probability of fraud and a 1 is the highest probability of fraud.

Score Category	Good Loans	Frauds
10	10	10
9	10	10
8	10	10
7	10	10
6	10	10
5	10	10
4	10	10
3	10	10
2	10	10
1	10	10
Total	100	100



Calculating the KS statistic

Score Cat.	Goods	Frauds	% Succ.	% Fail.	CDF Succ.	CDF Fail.	Diff.
10	10	10	10.0%	10.0%	10.0%	10.0%	0.0%
9	10	10	10.0%	10.0%	20.0%	20.0%	0.0%
8	10	10	10.0%	10.0%	30.0%	30.0%	0.0%
7	10	10	10.0%	10.0%	40.0%	40.0%	0.0%
6	10	10	10.0%	10.0%	50.0%	50.0%	0.0%
5	10	10	10.0%	10.0%	60.0%	60.0%	0.0%
4	10	10	10.0%	10.0%	70.0%	70.0%	0.0%
3	10	10	10.0%	10.0%	80.0%	80.0%	0.0%
2	10	10	10.0%	10.0%	90.0%	90.0%	0.0%
1	10	10	10.0%	10.0%	100.0%	100.0%	0.0%
Total	100	100	100.0%	100.0%	Maximum Difference=		0.0%

Yes this model is terrible, it doesn't predict anything.

KS statistic = 0.0%



Example of a Non Predictive Model

Below is another example of a non predictive model using more realistic numbers. This data assumes a 4% fraud rate.

Credit Scores Category	Total Number of Loans	Cumulative Percent	Number of Good Loans	Number of Frauds	Cumulative Percent Good Loans	Cumulative Percent Frauds	The Difference
10	200,000	10%	192,000	8,000	10.0%	10.0%	0.00%
9	200,000	20%	192,000	8,000	20.0%	20.0%	0.00%
8	200,000	30%	192,000	8,000	30.0%	30.0%	0.00%
7	200,000	40%	192,000	8,000	40.0%	40.0%	0.00%
6	200,000	50%	192,000	8,000	50.0%	50.0%	0.00%
5	200,000	60%	192,000	8,000	60.0%	60.0%	0.00%
4	200,000	70%	192,000	8,000	70.0%	70.0%	0.00%
3	200,000	80%	192,000	8,000	80.0%	80.0%	0.00%
2	200,000	90%	192,000	8,000	90.0%	90.0%	0.00%
1	200,000	100%	192,000	8,000	100.0%	100.0%	0.00%
Total	2,000,000		1,920,000	80,000	K-S statistic = Maximum Difference =		0.00%



The KS Statistic

Let us try again.

Does it look like this logistic model is predictive?

Score Category	Good loans	Frauds
10	31	0
9	25	1
8	17	2
7	10	5
6	6	6
5	5	6
4	3	7
3	2	10
2	1	20
1	0	43
Total	100	100



The KS Statistic

It does look predictive

Score Category	Good loans	Frauds
10	31	0
9	25	1
8	17	2
7	10	5
6	6	6
5	5	6
4	3	7
3	2	10
2	1	20
1	0	43
Total	100	100

A total 83 good loans out of 100, 83%, were placed into score categories 10-7.

A total 73 frauds out of 100, 73%, were placed into score categories 3-1.



Calculating the KS Statistic

Score Cat	Good	Fraud	% Succ.	% Fail.	CDF Succ.	CDF Fail.	Diff.
10	31	0	31.0%	0.0%	31.0%	0.0%	31.00%
9	25	1	25.0%	1.0%	56.0%	1.0%	55.00%
8	17	2	17.0%	2.0%	73.0%	3.0%	70.00%
7	10	5	10.0%	5.0%	83.0%	8.0%	75.00%
6	6	6	6.0%	6.0%	89.0%	14.0%	75.00%
5	5	6	5.0%	6.0%	94.0%	20.0%	74.00%
4	3	7	3.0%	7.0%	97.0%	27.0%	70.00%
3	2	10	2.0%	10.0%	99.0%	37.0%	62.00%
2	1	20	1.0%	20.0%	100.0%	57.0%	43.00%
1	0	43	0.0%	43.0%	100.0%	100.0%	0.00%
Total	100	100	100.0%	100.0%	Maximum Difference=		75.00%

This is a very good model. You can see this, you don't even need the KS statistic. This is an example of how important presentation is for understanding. The way in which we display the data allows us to quickly understand the predictive power of the model.

KS statistic
= 75.0%



How is it Applied?

Score Cat	Good	Fraud	% Succ.	% Fail.	CDF Succ.	CDF Fail.	Diff.
10	31	0	31.0%	0.0%	31.0%	0.0%	31.00%
9	25	1	25.0%	1.0%	56.0%	1.0%	55.00%
8	17	2	17.0%	2.0%	73.0%	3.0%	70.00%
7	10	5	10.0%	5.0%	83.0%	8.0%	75.00%
6	6	6	6.0%	6.0%	89.0%	14.0%	75.00%
5	5	6	5.0%	6.0%	94.0%	20.0%	74.00%
4	3	7	3.0%	7.0%	97.0%	27.0%	70.00%
3	2	10	2.0%	10.0%	99.0%	37.0%	62.00%
2	1	20	1.0%	20.0%	100.0%	57.0%	43.00%
1	0	43	0.0%	43.0%	100.0%	100.0%	0.00%
Total	100	100	100.0%	100.0%	Maximum Difference=		75.00%

Great, so now what? How do we apply this? Take a minute to think.



How is it Applied?

Score Cat	Good	Fraud	% Succ.	% Fail.	CDF Succ.	CDF Fail.	Diff.
10	31	0	31.0%	0.0%	31.0%	0.0%	31.00%
9	25	1	25.0%	1.0%	56.0%	1.0%	55.00%
8	17	2	17.0%	2.0%	73.0%	3.0%	70.00%
7	10	5	10.0%	5.0%	83.0%	8.0%	75.00%
6	6	6	6.0%	6.0%	89.0%	14.0%	75.00%
5	5	6	5.0%	6.0%	94.0%	20.0%	74.00%
4	3	7	3.0%	7.0%	97.0%	27.0%	70.00%
3	2	10	2.0%	10.0%	99.0%	37.0%	62.00%
2	1	20	1.0%	20.0%	100.0%	57.0%	43.00%
1	0	43	0.0%	43.0%	100.0%	100.0%	0.00%
Total	100	100	100.0%	100.0%	Maximum Difference=		75.00%

Imagine the rejecting of all loan applicants that are placed into score category 1. You would eliminate 43% of the frauds and not loose a single good loan.



How is it Applied?

Score Cat	Good	Fraud	Cum Succ.	Cum Fail.	Cum Odds
10	31	0	31	0	Inf
9	25	1	56	1	56.0
8	17	2	73	3	24.3
7	10	5	83	8	10.4
6	6	6	89	14	6.4
5	5	6	94	20	4.7
4	3	7	97	27	3.6
3	2	10	99	37	2.7
2	1	20	100	57	1.8
1	0	43	100	100	1.0
Total	100	100			

Interpret
Cum Odds

An Estimate of odds

$$\text{odds} = \frac{\# \text{ good}}{\# \text{ frauds}}$$

Think about what it means when we cut off the bottom 21.5% in terms of the odds and our example on fraud.



The Presentation

- A key to understanding is presentation. How do we view our results.
 - Visualization and presentation is very important.
- It is important to know your audience.
 - Your audience determines how you will present what you learn from the logistic regression model.
 - Senior management in a business is not interested in a theoretical data mining discussion. S/he is interested in how your fraud detection model will help the company.
 - A fellow statistician would need less visualization as they already understand, but in my opinion a nice presentation of results can only help.
- We will next cover how to look at the variables that enter into your model.
 - This is very important for gaining trust in your work.



How Do We View the Independent Variables in the Model?

- It is important to interpret the variable in the model and then look at the variable individually compared to the dependent variable.
 - Often the variable when viewed in the model might have the opposite relationship with the dependent variable than it does when looked at separately.
 - This can result from multicollinearity.
 - Multicollinearity will not be covered.
- Often when creating a model, it is good to think about the variables that enter into the model and why they are entered. You may be asked to explain why you choose to keep a certain variable and use it in the model.
 - One way to investigate the independent variable's relationship with the dependent variable is in the same way as when investigating the model.



Sample Partial Presentation Of A Fraud Detection Model

Included is only an explanation of
variables in the model and model
validation.



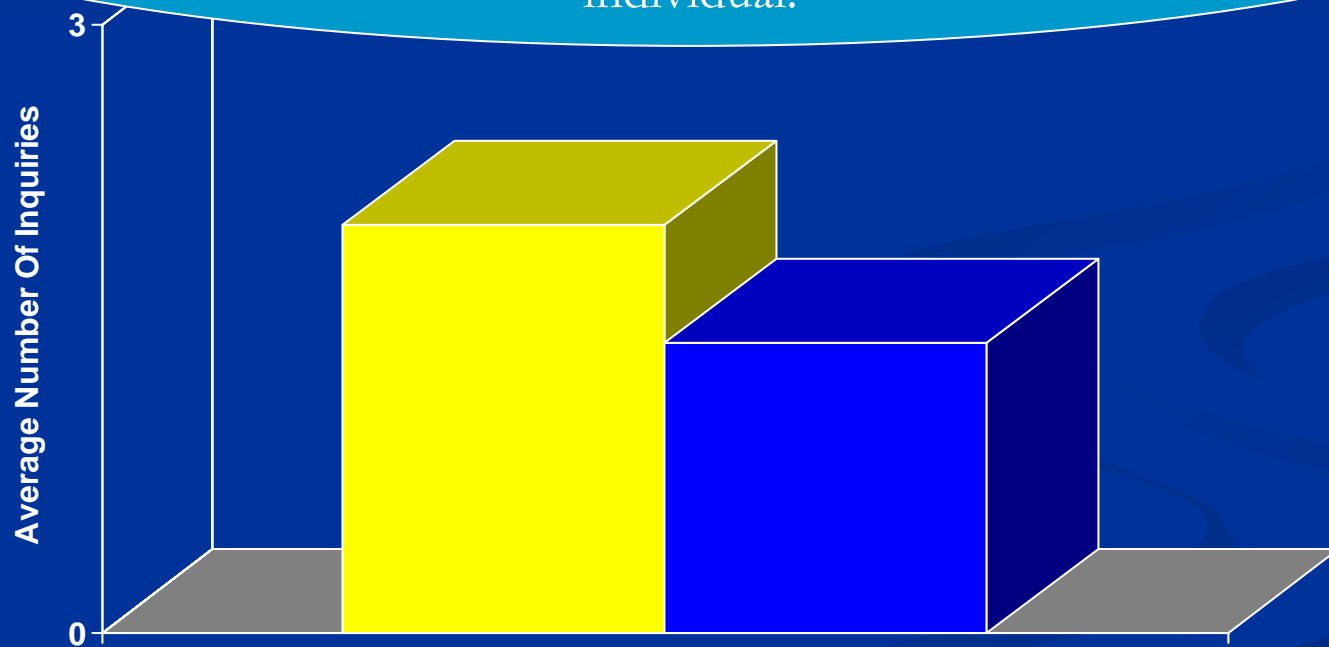
Most Important Factors For Detecting Fraud

Relationship With Fraud	The Variable
+	Number of inquiries for credit in the past 6 months
-	Driver License Number Match
-	Zip Code Match
-	Age of Applicant
-	Gender
Etc.	Etc.



Number Of Inquiries For Credit In The Past 6 Months

This slide is showing that people with more inquiries (applications) for credit are more likely to be a victim of fraud. Perhaps some of the inquiries for credit were made by someone attempting to commit fraud and not the actual individual.



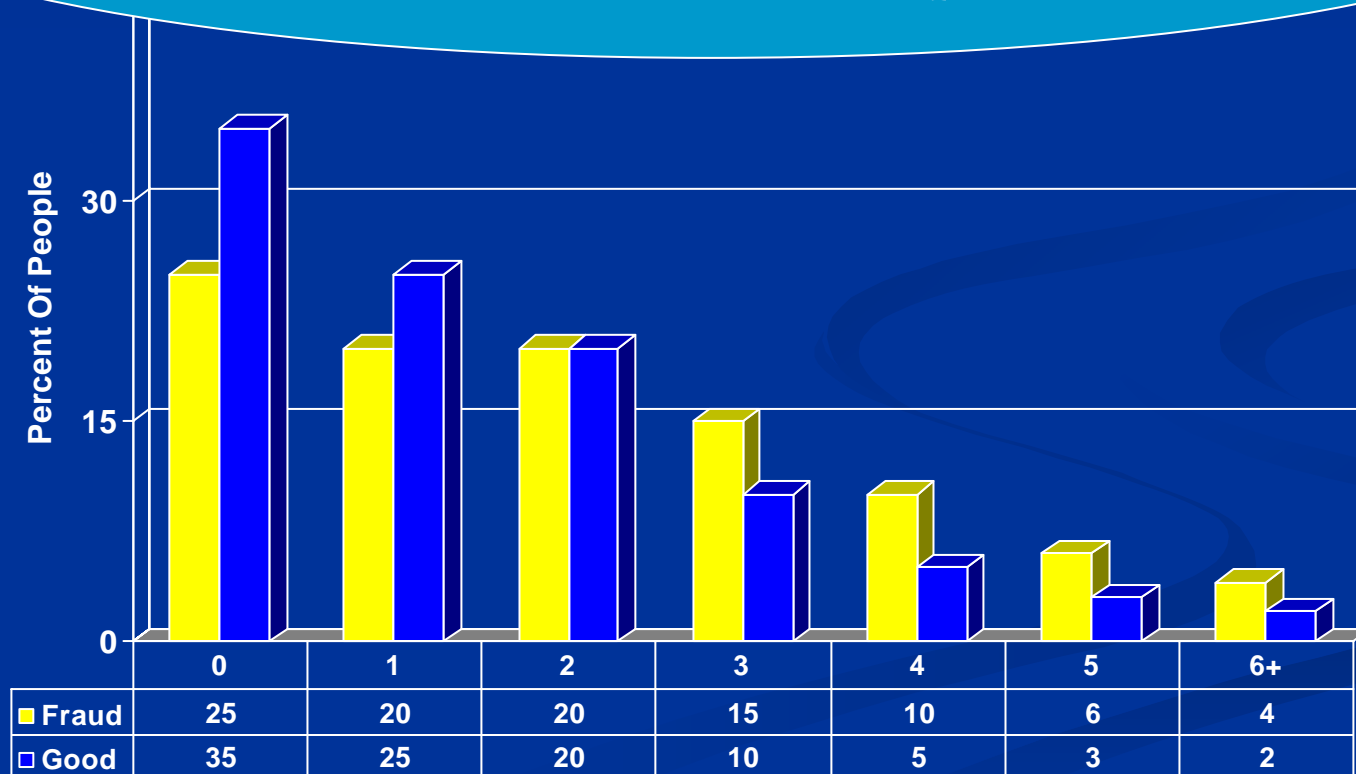
■ Fraud	2.01
■ Good	1.43



Number Of Inquiries For Credit In The Past 6 Months

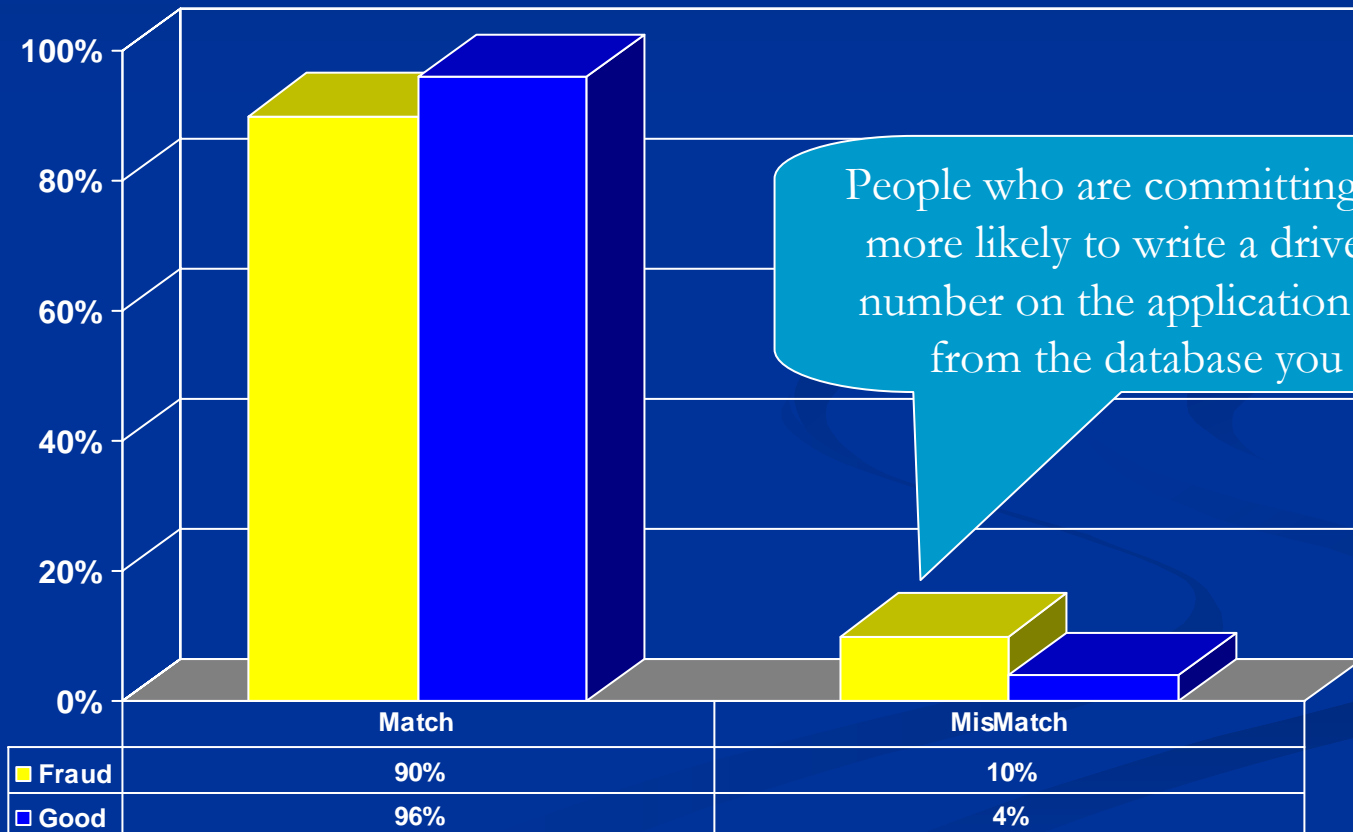
This slide is showing the same information as the previous slide. This slide is more informative, but many people will think the previous slide is better and easier to understand.

Know your audience (who you present to)!





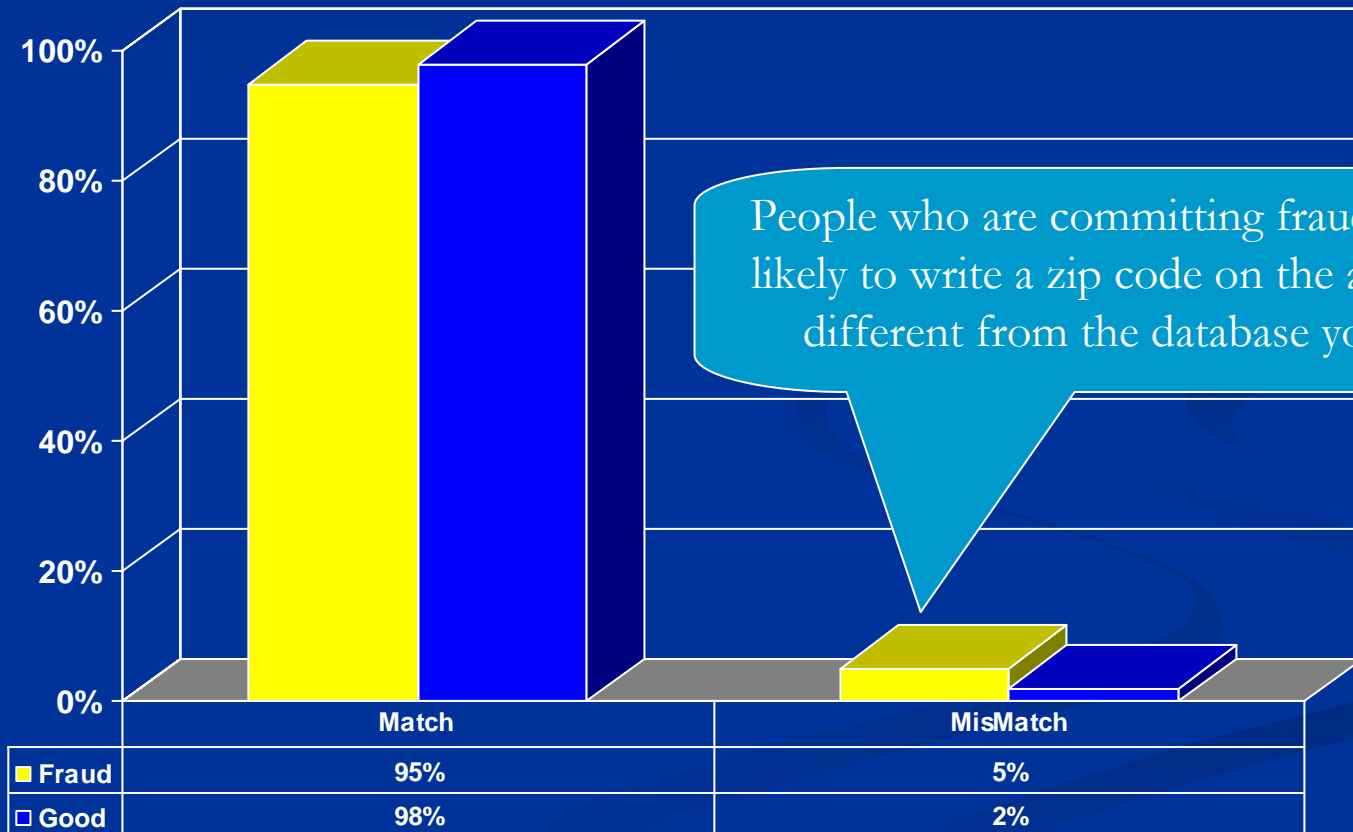
Percent Match and Mismatch Database On Driver License Number



People who are committing fraud are more likely to write a driver license number on the application different from the database you have.



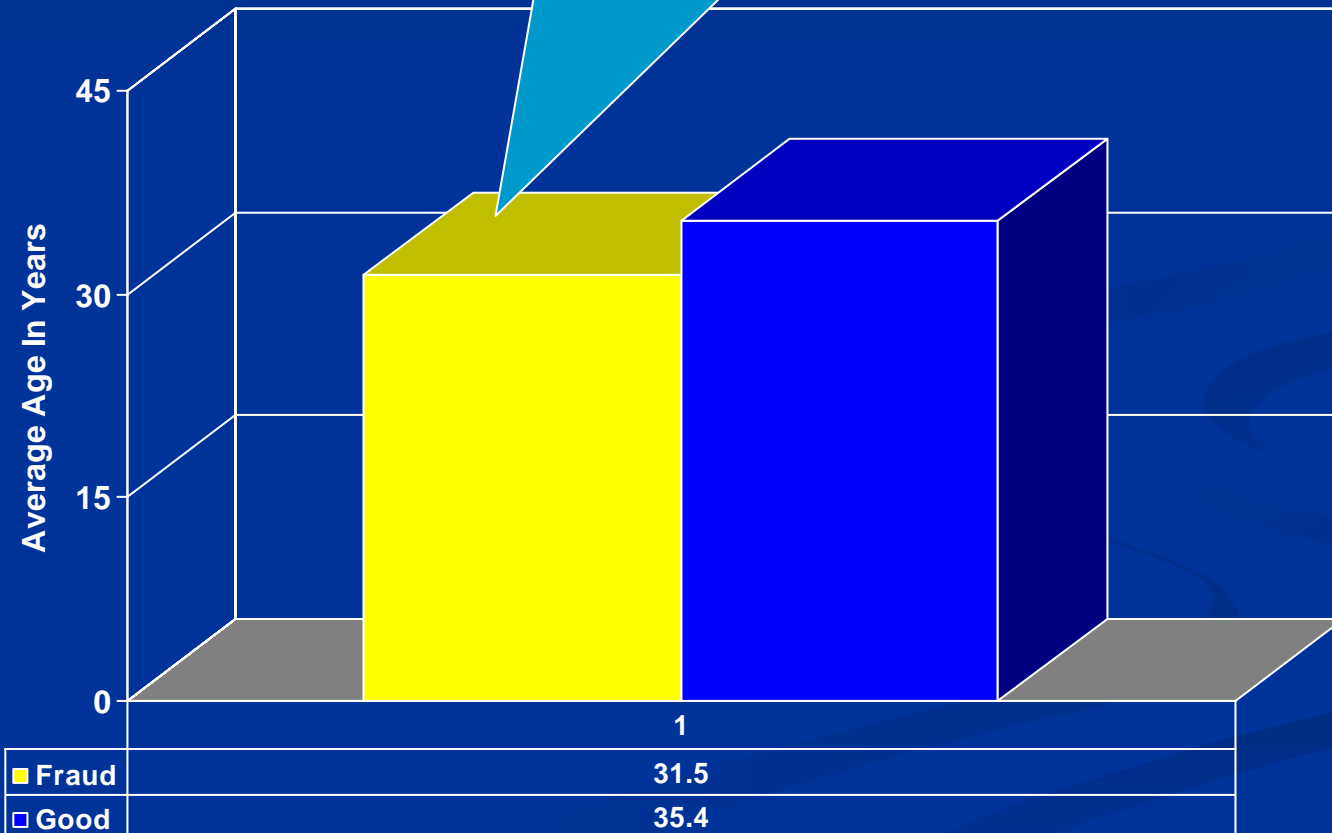
Percent Match and Mismatch Database On Zip Code





Average Age Of Applicant

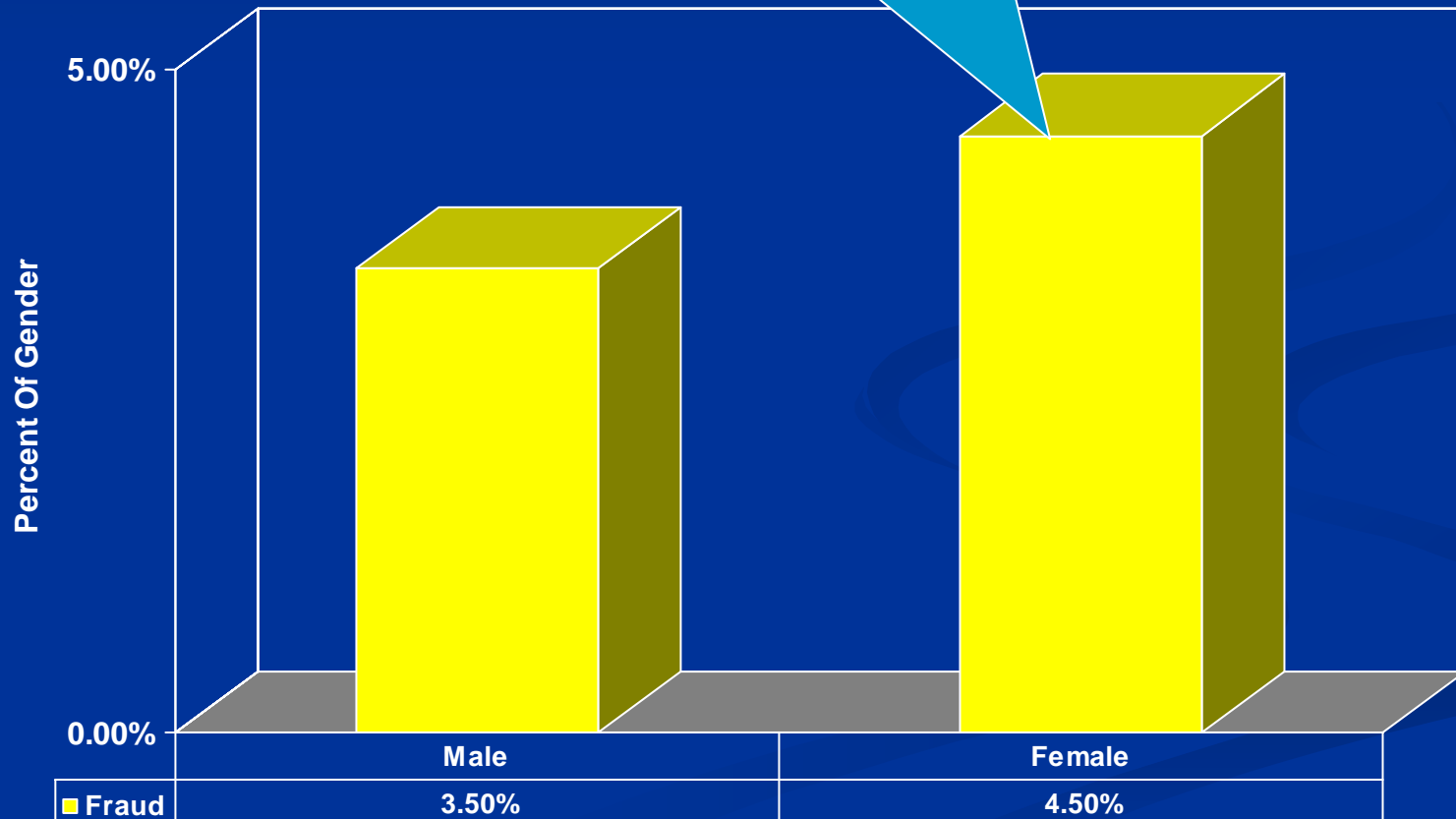
Younger people are more often victims of fraud.





Gender Of Applicant

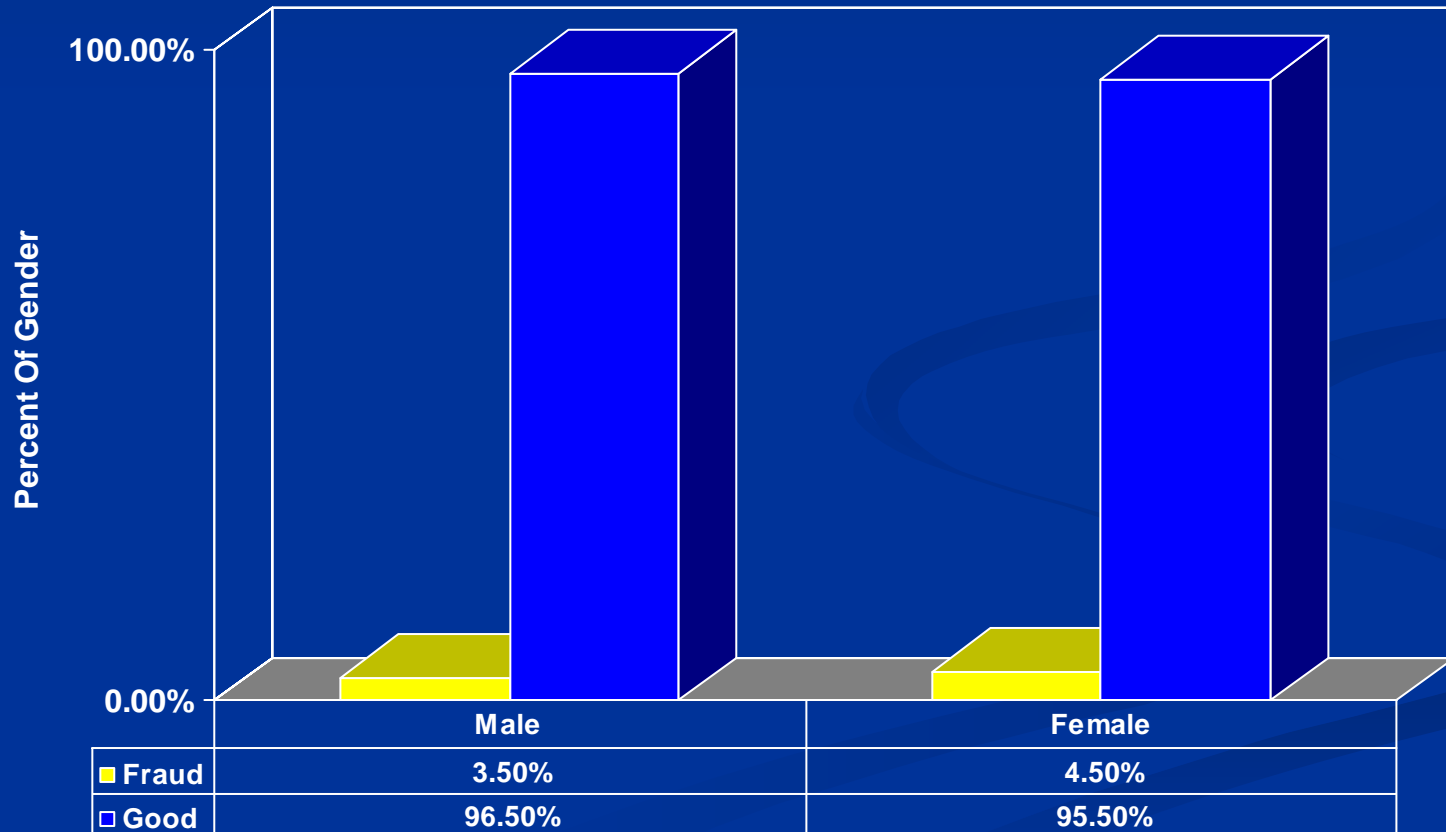
Females are more often victims of fraud.





Gender Of Applicant

Again there is more than one way to present the same thing.
Know your audience (who you present to)!





An More Graphs

- Those simple graphs would be produced for all variables in the model.



Understanding The Fraud Detection Model Performance

This Model has a KS of 25.82.

Credit Scores Category	Total Number of Loans	Cumulative Percent	Number of Good Loans	Number of Frauds	Cumulative Odds
10	200,000	10%	196,596	3,404	57.8
9	200,000	20%	196,170	3,830	54.3
8	200,000	30%	195,745	4,255	51.2
7	200,000	40%	194,894	5,106	47.2
6	200,000	50%	194,043	5,957	43.3
5	200,000	60%	193,617	6,383	40.5
4	200,000	70%	192,766	7,234	37.7
3	200,000	80%	191,489	8,511	34.8
2	200,000	90%	190,213	9,787	32.0
1	200,000	100%	174,468	25,532	24.0
Total	2,000,000		1,920,000	80,000	

By refusing the bottom 10% of applicants you can reduce fraud by 32% (25,532/80,000)

By refusing the bottom 10% you would have 32 good loans to one fraud, before 24 good loans to one fraud.



Concluding Remarks

The Data

The Technique

The Presentation

When all three are combined properly you have a very powerful data mining and data analysis tool in general.