**Modeling Strategy for Expeditors Risk Management**

**Business Context:**

For every shipment there is an inherent risk of financial loss resulting from a claim. At a minimum, our customers receive standard insurance coverage to protect against a partial loss if the full value is larger than some threshold. If the value is less than the threshold then our customers are fully covered. When our customers are not fully covered they can request extended coverage. With extended coverage, if a loss occurs, our customers will be compensated fully; however, extended coverage negates our insurance thereby transferring the full liability risk onto us. Therefore, it is critical to understand the cost of risk for decision making.

*Future considerations:* Insurance is costly. When we utilize insurance we pay not only for expected losses but we also pay for the insurance company’s cost of doing (risk loading). Given that we can financially afford to pay for claims out of pocket, we should self-fund our insurance services to save on the administrative costs. Two scenarios below are provided to outline this line of thinking.

If we expect to profit from insurance then we are saying that the insurance companies are losing money on us (zero sum game). If we don’t expect insurance companies to be losing money on us, to what extent are they profiting from us? If the cost of evaluating such benefits outweighs the expected benefits then we should abandon such prospects. Currently, we are abandoning such a cost-savings prospect.

**Abstract:**

The purpose of this outline is to provide an understanding of how we can model our liability risk, the tools that will be developed to utilize the models, the process we take in determining which variables to model, how we define those variables, and what methods we will take to maximize the information gained from the modeling process. In insurance, the liability cost is referred to as the “pure premium” or “expected loss”, henceforth refered to as the “cost of risk” or “liability per shipment”, as it only takes into account the risk of loss and not the administrative costs associated with managing claims. The components of the model include the probability of a claim occurring, the expected dollar amount paid to the claimant given a claim occurs and the expected dollar (recovery) received from the carrier:

**Cost of Risk = P(Claim) \* E(Paid to Claimant|Claim) \* E(Recovery Rate|Claim)**

Two other measures of risk we will consider includes unexpected loss and value-at-risk (also tail value at risk). Unexpected loss is defined as the standard deviation of expected loss. You can find the formula for that page 4. Value at risk is the measure of loss such that x% of the time you can expect to have a loss of **at least** $Y. For example, if you compute a 99% VaR then you would say that 99% of the time your loss will be at most $Y which if you say compute a 1% VaR then you would say that 1% of the time you can expect to lose at least $Y. Tail value at risk (TVaR) can be interpreted as the expected value of loss given that a loss of at least amount $Y occurs.

In order to make better decisions involving variables/parameters of shipments and comparison of liability terms we will develop models to calculate each of the described measures of risk.

**Fundamental Risk Measure: Cost of Risk – Valuation Approach**

Although each component of the modeling process is important (probability of claim, expected recovery and loss) the pure premium is the single value that is utilized in determining whether break-even, within the respective shipping process, at a minimum has been achieved. This pure premium is calculated as the probability of a claim, multiplied by the expected loss size should a loss occur, multiplied by the expected carrier recovery rate. The other components will be modeled with corresponding sensitivity outlined but the following is an example as to why we utilize the pure premium as the single measure of the risk associated with the business being quoted:

**Example I: Two distinct shipments, same expected loss size of $100,000**

* Company A ships from SHA to MIA and has a probability of claim per shipment of 0.01%
* Company B ships from ORD to MIA and has a probability of claim per shipment of 0.005%
* 🡺 Disregarding recovery, company A has an expected cost of risk of $100,000 X 0.01% = $10
* 🡺 Disregarding recovery, company B has an expected cost of risk of $100,000 X 0.005% = $5
  + Although expected loss is the same for both companies, after repeated shipping a premium charged of $10 will result in no profit from doing business with company A while the same premium would result in profit from company B equaling the # of shipments multiplied by $5

Now, we do know that variability will certainly factor into the equation of value to some extent, but those measures will mostly help to determine which risks will be nonetheless acceptable.

**Risk Calculator Features: Valuation & Value-at-Risk Approach**

Two methods should be considered in light of risk management, valuation and value-at-risk. Valuation utilized expected future cash flows and then discounts back to the present for cost purposes. Value-at-risk utilized future cash flow at a stated confidence level, say 5%. A 5% VaR is an amount such that there is a 5% chance that losses will be at least $X. One of the end goals of this project is to develop a calculator to estimate risk metrics for aggregated shipments (transactional on a limited basis). Sensitivity and scenario analysis will be possible with the output from these models and thus accounts for the VaR part of this exercise. Some features of this calculator to display are as follows:

* Estimated pure premium on a per shipment and aggregated basis (need to determine time span)
  + Standard and Extended, cost difference
* Estimated cost per loss on a per shipment and aggregated basis
  + Standard and Extended, cost difference
* Estimated recovery rate on a per shipment and aggregated basis
  + Standard and Extended, rate difference
* Scenario analysis and sensitivity analysis surrounding operational risk and model risk
  + Standard and Extended, different recovery rates, different fixed extended terms
* Measures based on VaR and TVaR
  + Standard and Extended
* Unexpected Loss (one number)
  + Standard and Extended

**Recovery Rates: Standard vs Extended**

In the following section, the distinction is made between two types of so-called recovery rates. The first one is in relation to the scenario where Expeditors had insurance coverage. When coverage is provided, the max payout for Expeditors is $10K. If the carrier is at fault to some extent they will pay their appropriate share. If that amount is greater than the amount the insurer covered then Expeditors will receive some “recovery”, in which the recovery rate will be expressed as a fraction of the $10K deductible. However, if no coverage were provided, then the recovery rate should be defined as the amount the carrier paid as a fraction of the total paid to the claimant. Thus two distinct “recovery” rates need to be modeled in order to estimate Expeditors liability under both scenarios, otherwise an arbitrary (or an incorrect) recovery rate will be applied and the cost of risk will be misquoted.

**Example II: Paid to claimant is $100,000. Expeditors deductible is $10,000**

* **Scenario 1: Carrier admits 95% fault (hypothetical):**
  + **🡺** Insurance receives $90,000 back and Expeditors receives $5,000 back
  + 🡺 Standard EI Recovery Rate of $5,000 / $10,000 = 50%
  + 🡺 If Expeditors had made this extended coverage fixed at $100,000
    - Expeditors recovery rate would have been $95,000 / $100,000 = 95%
* **Scenario 2: Carrier admits 80% fault (hypothetical):**
  + 🡺 Insurance receives $80,000 back and Expeditors receives $0 back
  + 🡺 Standard EI Recovery Rate of $0 / $10,000 = 0%
  + 🡺 If Expeditors had made this extended coverage fixed at $100,000
    - Expeditors recovery rate would have been $80,000 / $100,000 = 80%
* ***Note:*** Upon viewing our historical data, we will observe extended recovery rates based on the observed data from situations where we had standard liability terms. It is believed that these values will be understated as we are more likely to recover a larger amount from the carrier in the event that we have more at stake.
* Therefore, in order to provide accurate predictions in going from standard to extended terms, we need to have two different recovery rates to account for both types of scenarios and the effect of priority of the party who recovers.

**Claims with 100% Carrier Fault**

There is an argument for excluding claims that received full recovery from the data set. The rationale is as follows: We were almost certain at the time of incident that the carrier was at fault and we weren’t liable. We nevertheless had to file the claim on behalf of the customer. Therefore we should exclude the claim from the data set and only model based on claims that Expeditors paid.

The opposing argument is that on a forward looking basis we have no way to know if a particular claim will result in Expeditors making a net payment to the customer. It is also believed that modeling under either scenario will result in identical (or nearly identical) results. The key differentiator of positions is that the customer doesn’t care whether it is Expeditors who is liable or the carrier. They only care about the risk of loss. Thus, we shouldn’t exclude the data under consideration since the results won’t change and our customers will have a richer amount of information regarding their shipments.

**Stratification Methodology:**

From the book “Applied Analytics” beginning on page 6-29 reads, “*A common predictive modeling practice is to build models from a sample with a primary outcome proportion different from the true population proportion. This is typically done when the ratio of primary to secondary outcomes is small.”* Further*, “The advantage of separate sampling is that you are able to abtain (on the average) a model of similar predictive power with a small overall case count. This is in concordance with the idea that the amount of information in a data set with a categorical outcome is determined not by the total number of cases in the data set itself, but instead by the number of cases in the rarest outcome category. (For binary target data sets, this is usually the primary outcome.) (Harrel 2006)*. One main reason for this practice is that without using separate sampling, a good model would be built by simply classifying everything as a non-responder (accuracy will be (1-response rate) which is approximately 99.9% in this project). In the beginning of the modeling exercise, we will specify a “prior” probability that represents the overall claim rate based on the entire population and SAS EM will convert all scores back to the valid probability measure. If we utilize another program then we will simply back-transform the P(Claim) to its appropriate value.

Next, this leads us to create a stratified random sample. The idea here is to have the total number of shipments equal to the total number of claims multiplied by five such that 20% of the shipments represents claim. The stratification will be based on having a proportional number of random non-claim shipments to each respective month represented in the modeling data set. This means that if a particular month happened to have significantly less total shipments than any of the other months then the total random samples selected from that month will be proportionally less.

This procedure can be done in Base SAS via the SURVEYSELECT proc statement with the stratification statement. For more on this procedure, consult the Advanced Analytics folder named, “SAS Code”.

**Modeling**

This liability modeling exercise takes into account three separate models with a few imposed output constraints depending on which case we are considering to accomodate business sense: (1) the probability of a shipment resulting in a claim, (2) Expeditors net liability recovery rate from the carrier, and (3) the full value paid to the claimant. Full value paid to the claimant is roughly Expeditors net liability plus the carrier’s net liability plus Insurance net liability.

**Modeling Option 1:**

For Pr(Claim), models will be developed by each mode for each loss type. For the expected paid to claimant, E(L), there will be one model built for each loss type within each mode. For the expected recovery rate model E(RR) there will be a model built for standard / extended within each loss type within each mode. Thus, there will be a total of 6 + 6 + 12 = 24 models built.

**Modeling Option 2 (preferred):**

Under this approach we will be able to identify process points and reduce the number of models significantly and provide added information for process improvement. The idea here is to incorporate an added set of variables to indicate whether particular process points took place or not.

Expected Loss:

**Unexpected Loss:**

**Standard & Extended**

**Unexpected Loss *(Standard Deviation of aggregate loss) =***

**Modeling Justification:**

Several types of specific models are proposed below under the header “*Modeling Methods*”. Two specific outcomes are of interest: expected outcomes and unlikely outcomes. The expected outcomes are addressed below and should be familiar with most. This is just the result of most regression-like modeling procedures. The unexpected outcomes may be less intuitive for most analysts and even modelers. The first option is the quantile regression method, utilized in Base SAS and R. With these types of models we can focus on more extreme quantiles of the distribution versus utilizing sensitivity analysis *(see appendix for more detail).* This enables us to model value-at-risk and tail value at risk. Another interesting model is the Loglinear Variance Model which will model the variance of outcomes opposed to the expected value of an outcome. With this method we will be able to model the variance of the expected recovery rate which will enable us to utilize the unexpected loss formula described above. Each method has its pros and cons which is why we will incorporate both.

Based on the Unexpected Loss formula, the only extra model that needs to be generated is a variance model for the Recovery Rate variable. The reason for this is straightforward. First, the expected claim frequency is simply the number of shipments multiplied by the probability of a claim. Thus, all we need to add here is the number of shipments which are predetermined. Second, the variance of the claim frequency is simply the variance formula for a binomial distribution (n \* p \* (1-P)). This leaves two inputs: expected recovery and variance of recovery. The expected recovery will already be modeled. However, the variance of the recovery needs to be modeled and this is where the Loglinear Variance model becomes useful. If the Loglinear Variance model fails to converge, we will simply use a binomial approximation (again, n \* p \* (1-p)). This will likely not skew the results too much as the recovery rate’s range is [0,1].

Modeling expected loss from an insurance standpoint is straightforward and has been conducted in similar ways for a long time. The below exerpt from the book, “Generalized Linear Models for Insurance Data”, explains this traditional method.

*“The models for claim size in Chapter 8 are for positive claim sizes: they condition on a claim having been made. Analysis of claim size traditionally treats the occurrence of a claim and the amount of a claim, given that there is a claim, as separate models. The former is usually a logistic regression and the latter a model such as a gamma or inverse Gaussian regression. The canonical link for the gamma distribution is the inverse function.* *The results of these separate models are combined to produce predictions of expected claim sizes.” Since parameters from a model with inverse link are difficult to interpret, the log link is usually regarded as more useful.*

*“Continuous responses of interest to insurers include claim size and time between the reporting of a claim and settlement. Continuous insurance variables**are usually non-negative and skewed to the right. Options for modeling**these variables are:*

*• Use a transformation to normality, and then employ the normal linear model on the transformed response. Thus g(y) ∼ N(μ,σ2) where g is the transformation and μ = xβ. The normal model is dealt with in Chapter 4.*

*• Generalized linear modeling, using a response distribution that is concentrated on the non-negative axis. Examples are the gamma and inverse Gaussian distributions.*

**Modeling Methods:**

* **Imputation / Transformation:**
  + Categorical Inputs:
    - Group Rare Levels: Identify top 20 – 30 levels
      * Branch Origin Code
      * Branch Destination Code
      * Carrier
      * Customer
  + Continuous Inputs:
    - Best possible transformation: Continuous Inputs
      * Maximum correlation
      * Standardize
      * Optimal Binning
  + Continous Target: E(Paid to Claimant)
    - Log or Standardize

**Models:**

* **Pr(Claim) Models:**
  + Enterprise Miner Models:
    - Air Missing, best to worsts, based on ROC:
      * Dmine Regression
      * Neural Network
      * Decision Tree
      * Logistic Regression
      * DMNeural
      * Gradient Boosting
    - Air Damage, best to worst:
      * Decision Tree
      * Logistic regression
      * Dmine Regression
      * DMNeural
      * Support Vector Machine
      * Gradient Boosting
* **Expected Paid to Claimant E(Full Payout | Claim):**
  + Enterprise Miner Models:
    - Air Missing, best to worst, based on EM selection
      * Neural Network
      * Partial Least Squares (PLS)
      * DMNeural
      * Least Angle Regression (LARS)
      * Regression
      * Dmine Regression
    - Air Damage (best to worst)
      * DMNeural
      * Regression
      * Partial Least Squares
      * Least Angle Regression
      * Neural Network
      * Dmine Regression
      * Quantile Regression (round 2) (Base SAS, R): *see Appendix A*
* **Expected Recovery Rate**
  + Enterprise Miner Models
    - Air Missing Standard:
      * Gradient Boosting
      * Neural Network
      * Decision Tree
      * Dmine Regression
      * Regression
    - Air Missing Extended:
      * Gradient Boosting
      * Neural Network
      * Dmine Regression
      * Decision Tree
      * Regression
    - Air Damage Standard:
      * Gradient Boosting
      * Neural Network
      * Dmine Regression
      * Decision Tree
      * Regression
    - Air Damage Extended:
      * Neural Network
      * Gradient Boosting
      * Dmine Regression
      * Decision Tree
      * Regression

**Model Validation**

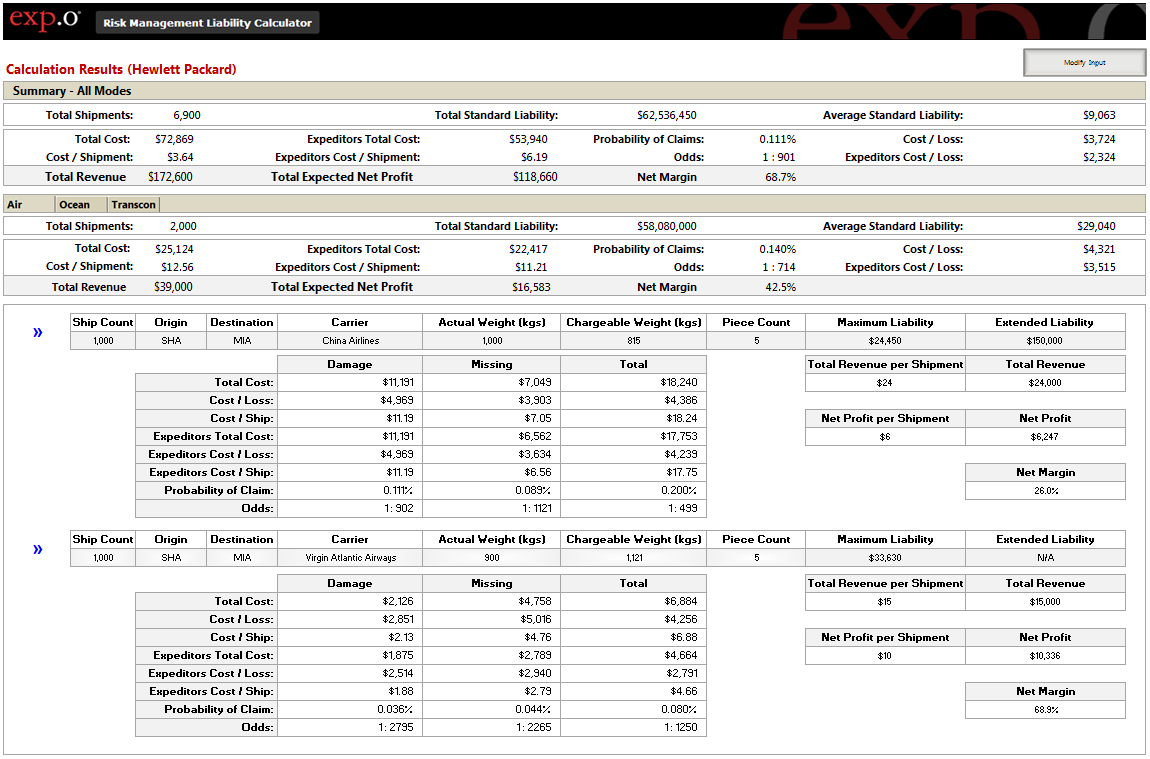
Typical model validation includes calculating various loss measures such as MAPE. This is not sufficient, however, for this application. We are charging a premium for each shipment and funds accumulate. Over time, the fund diminishes at discrete points in time as claims occur. The main question is this: Are we charging an appropriate premium? A secondary question of interest is this: Are we appropriately discriminating our segements by charging a higher premium to some and lower premium to others?

The preceding paragraph naturally leads to a desirable validation method. First, we compare the accumulated premiums to the aggregate claims paid. Secondly, and a bit more labor intensive, we do the same analysis for each discriminated segement.

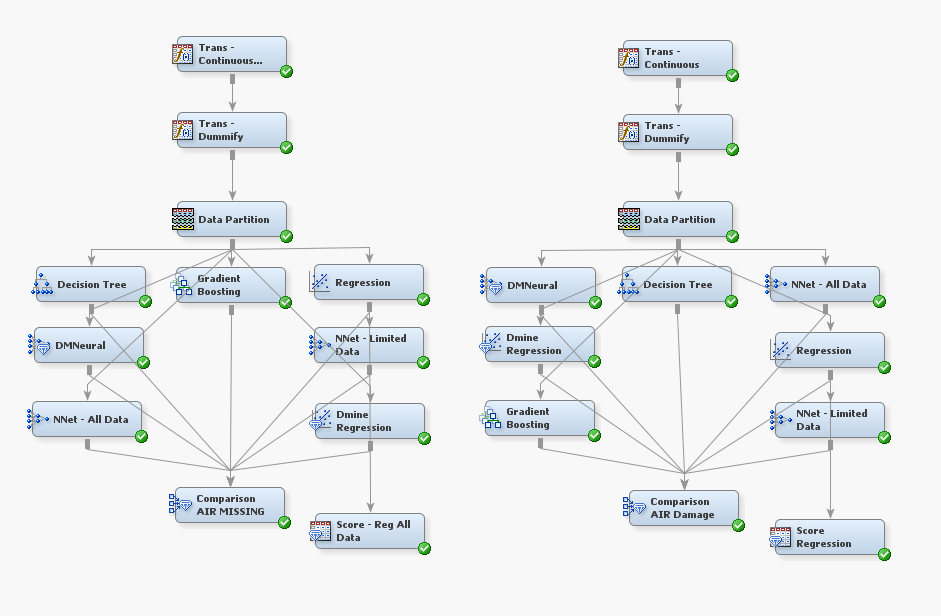
Another natural question is this: How much time is needed so that we can be confident at some stated level that we are in fact charging the correct amount? This question is essentially a means to understand at what point to we go from short-run to long-run in a ‘law of large numbers’ context.

**APPENDICES:**

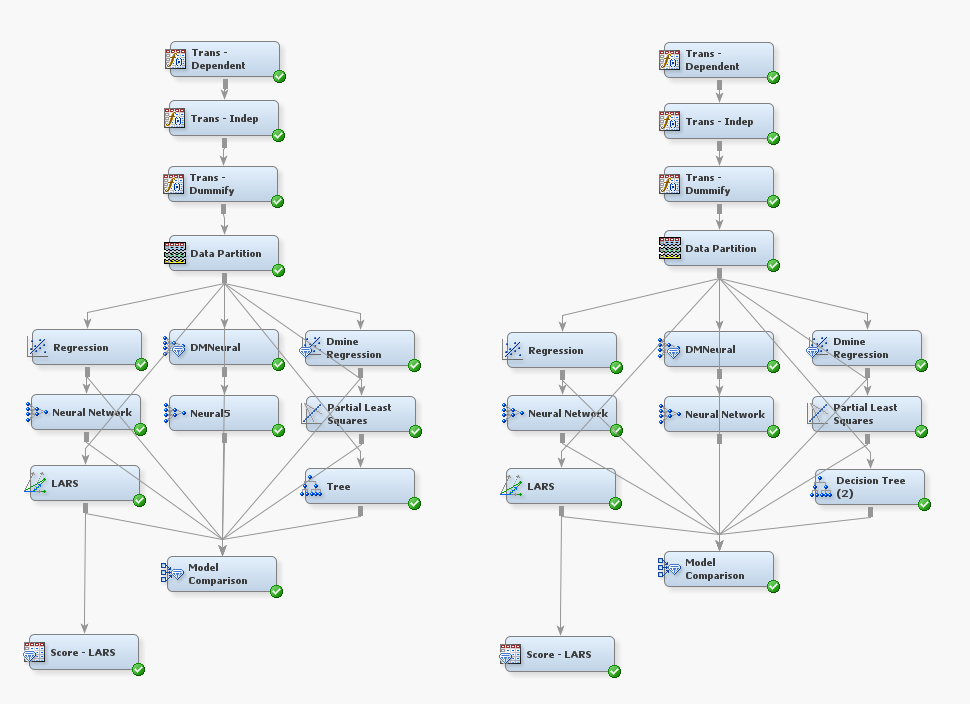
**User Interface of Application:**



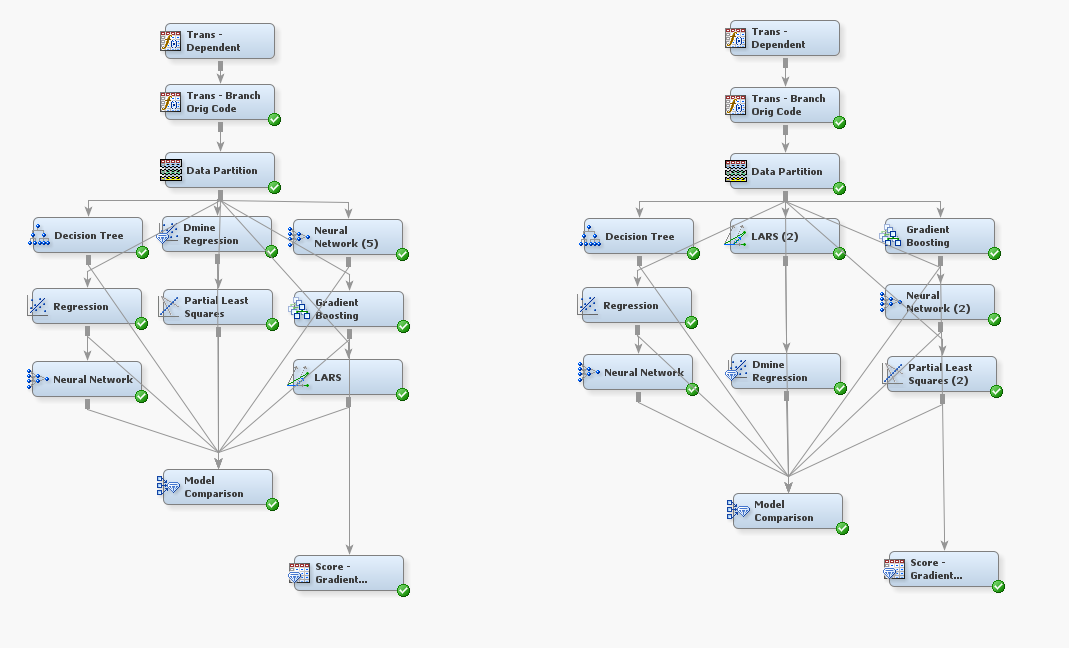
**Probability of Claim: Missing and Damage Models**



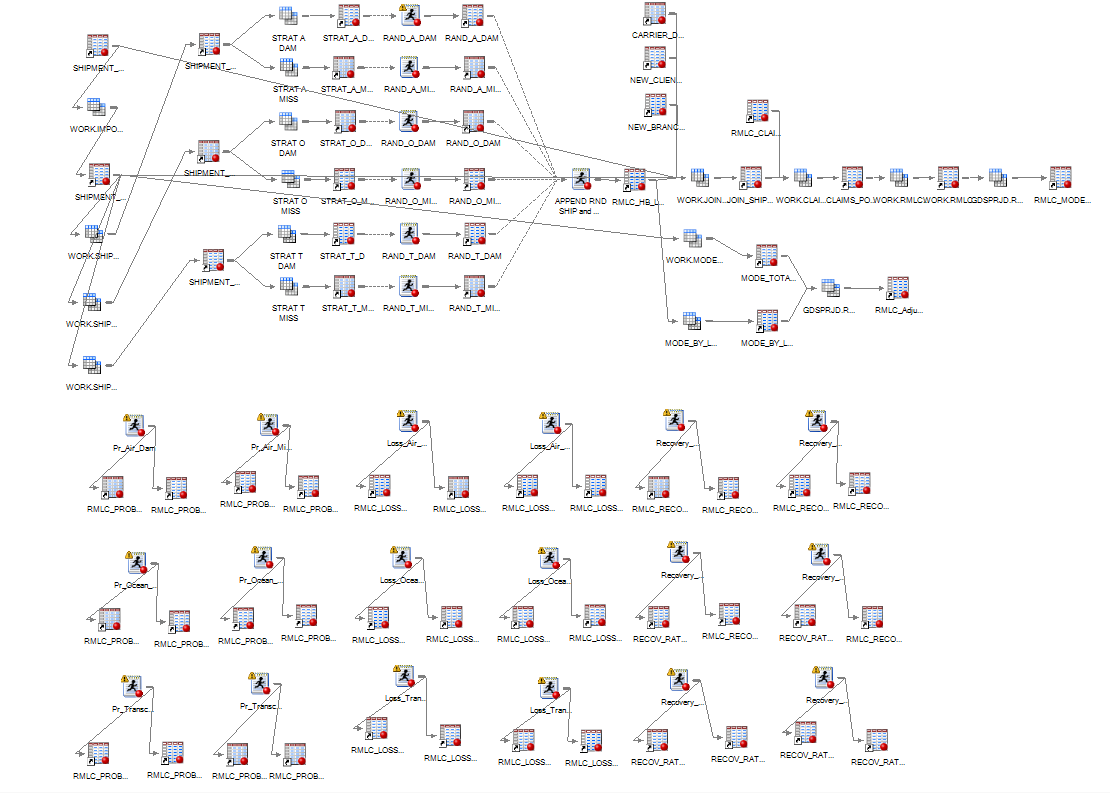
**% Loss of Max Liability Models:**



**Recovery Rate Models:**



**DATA PREP:**



**Utilizing the Quantile Regression Procedure, VaR and TVaR Meaures:**

Quantile regression is similar to other modeling tools. Other modeling tools predict expected outcomes while quantile regression predicts quantiles, hence the name. The reason for this modeling consideration is that in the context of risk management, often times the manager is concerned with adverse outcomes. While there are many basic measures used for risk analysis, a more advanced method includes VaR (value at risk) or TVaR. VaR is nothing more than a quantile from a distribution relating a probability of occurrence and a dollar cost associated with this occurrence. TVaR is the average of all losses that fall below the VaR. In this modeling project we will be focusing on the VaR approach and consider modeling TVaR. It should be noted that TVaR is the preferred method for analyzing risk for several reasons:

* 1. It satisfies all the coherent risk measure: Monotonicity, subadditivity, positive homogeneity, and translation invariance
  2. It provides an estimate of the magnitude of a loss for unfavorable events conditional on these unfavorable events occurring while VaR only tells you what the loss would be at a stated level of confidence

One way to model a VaR measure is to build a regression model that predicts quantiles of interest, hence potential utilization of the quantile regression procedure. One way to model TVaR is to set up a set of quantile regressions focuses solely on the quantiles less than some confidence level and to average the outcomes of such predictions.

Below is a chart representing an example of 10 fitted quantile regressions and their predicted responses over a covariate space. As can be seen, there is great benefit for describing the nature of potential risks with models like these. Based on the below graph we can say that for low or high values of the x-variable that there is less dispersion (risk) whereas in the middle of the distribution there is greater dispersion (risk).

