Advanced Analytics Consulting Services, LLC
(AA Consulting)

2019 Conference on Statistical Practice
February 16, 2019

Targeting Return-to-Work Intervention by Predicting Prolonged Workers' Compensation Claims

Mei Yu Najim, CSPA, Advanced Analytics Consultant
Advanced Analytics Consulting Services, LLC

Mei Yu Najim, CSPA
Advanced Analytics Consulting Services, LLC
(AA Consulting)

Mrs. Mei Yu Najim provides advanced analytics consulting services including developing full life cycle predictive modeling processes from raw data exploration to model implementation into IT data systems and providing thorough documentation and related training to the P&C insurance industry.

Mei has 15 years hands-on advanced analytics experience including statistical methods and machine learning algorithms (GLM, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Neural Network, etc.) and extensive data mining experience dealing with large and complex data sets in various predictive analytics settings (claims, underwriting, pricing). She also has experience in cat modeling, actuarial pricing, reserving, and R&D. She has frequently presented at conferences to share her expertise in predictive analytics.

Mei holds a Bachelor of Science in Actuarial Science from Hunan University and two Master of Science degrees, in Applied Mathematics and in Statistics, from Washington State University. Mei is a member of the American Statistical Association and a Certified Specialist in Predictive Analytics (CSPA) of the Casualty Actuarial Society.

Advanced Analytics Consulting Services, LLC
(AA Consulting)

AGENDA

- Introduction to Advanced Analytics and Machine Learning
- Insurance Claims Analytics
- Return-to-Work Day 30 Model
- Q & A

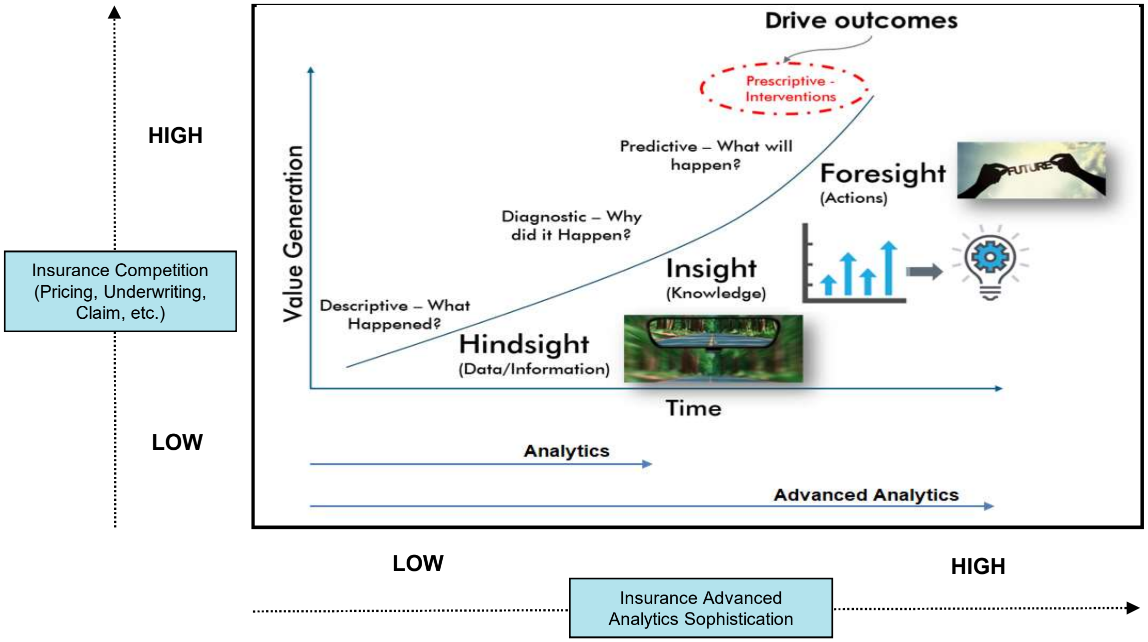
Advanced Analytics Consulting Services, LLC
(AA Consulting)

AGENDA

- Introduction to Advanced Analytics and Machine Learning
- Insurance Claims Analytics
- Return-to-Work Day 30 Model
- Q & A

Advanced Analytics Consulting Services, LLC
(AA Consulting)

Different Level of Analytics


HIGH

LOW

Insurance Competition (Pricing, Underwriting, Claim, etc.)

LOW

HIGH

Insurance Advanced Analytics Sophistication

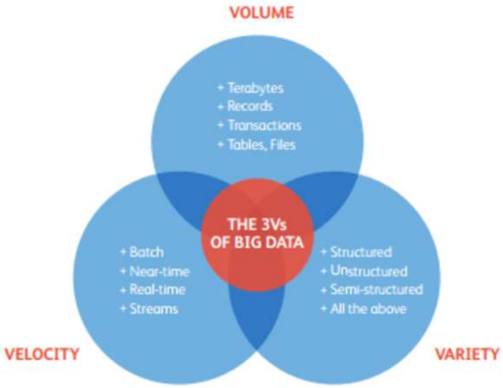
Note: In the P&C Insurance industry, the more advanced analytics sophistication a company has, the more competitive it can be.

- Introduction to Advanced Analytics and Machine Learning

Traditional Data vs. Big Data

Big Data (Volume, Variety, and Velocity) → Big Data Strategy (Value)

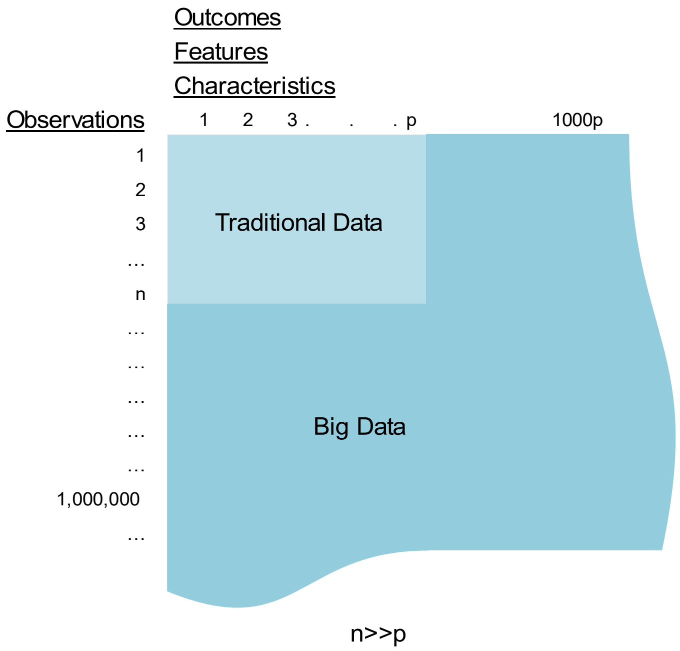
- Volume: MB→ GB→ TB → PB
- Variety: Structure, Semi-structure, Un-structure, Photo, Web, Audio, Social, Video, Mobile
- Velocity: Batch → Periodic → Near Real Time → Real Time → Streams



- Introduction to Advanced Analytics and Machine Learning

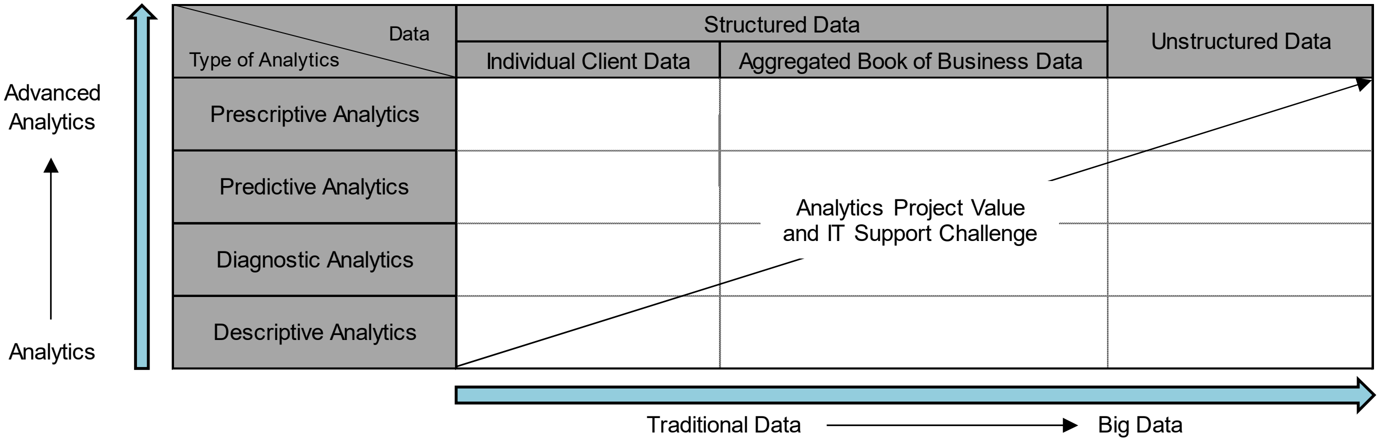
Traditional Data vs. Big Data

Data Volume: MB→ GB→ TB → PB (millions of records and thousands of data fields)


Outcomes
Features
Characteristics
Observations
1 2 3 . . . p
1000p
1
2
3
…
n
…
…
…
…
…
1,000,000
…
Traditional Data
Big Data
n>>p

- ➢ Introduction to Advanced Analytics and Machine Learning

Analytics Project Value and IT Support Challenge



Type of Analytics / Data	Structured Data		Unstructured Data
	Individual Client Data	Aggregated Book of Business Data	
Prescriptive Analytics			
Predictive Analytics			
Diagnostic Analytics			
Descriptive Analytics			

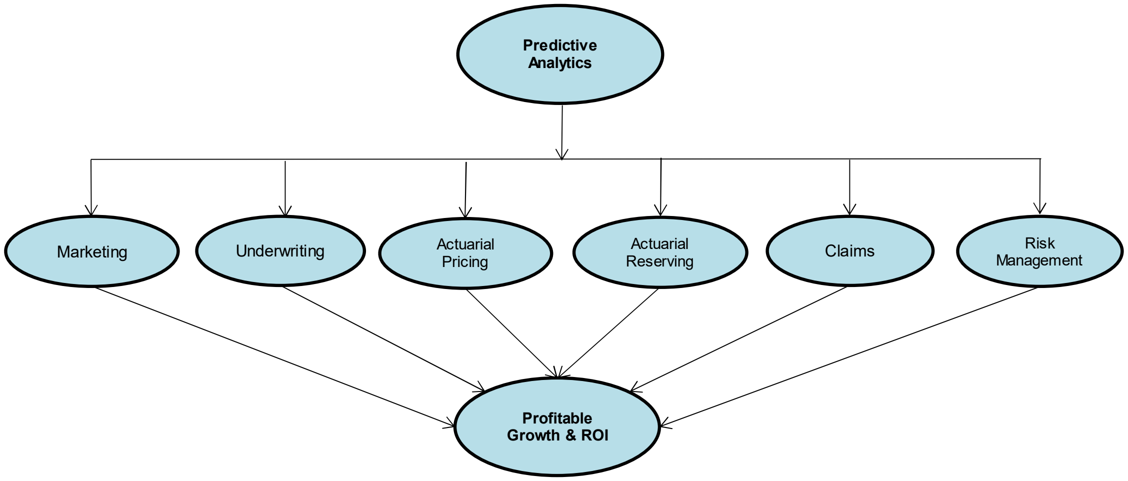
Advanced Analytics Consulting Services, LLC
(AA Consulting)

AGENDA

- ➢ Introduction to Advanced Analytics and Machine Learning
- ➢ Insurance Claims Analytics
- ➢ Return-to-Work Day 30 Model
- ➢ Q & A

Advanced Analytics Consulting Services, LLC
(AA Consulting)

Core Operations Using Predictive Analytics in P&C Insurance



Note: In the Property & Casualty Insurance industry, advanced analytics has increasingly penetrated into each of the core business operations – marketing, underwriting, actuarial pricing, actuarial reserving, claims, and risk management, etc. Advanced analytics is becoming one of the important ways to grow revenue and increase profit in the insurance industry.

Advanced Analytics Consulting Services, LLC
(AA Consulting)

An Overview of Claims Analytics

Operation	Now*		Two Years Forecasting	
	Personal	Commercial	Personal	Commercial
Claim triage	18%	15%	59%	66%
Fraud potential	28%	14%	70%	55%
Litigation potential	23%	10%	54%	50%
Report ordering	34%	17%	74%	48%
Case reserving	9%	8%	41%	48%
Loss Control	N/A	2%	N/A	39%
Marketing and advertising	21%	N/A	39%	N/A

* Survey fielded September 7 - October 24, 2016 and released March 2017
Source: Towers Watson – 2016 Predictive Modeling Benchmark Survey (U.S.)

Notes:
1. Use of predictive modeling in core operational areas of insurance companies is projected to grow substantially in the coming years.
2. Claims analytics is mainly about better claim management and assisting adjusters in making good, consistent decisions about how to handle a given claim

- Insurance Claims Analytics

Some Common WC Claim Predictive Models

Complex or Large Loss Model predicts the claim overall severity to identify complex or large loss claims above certain thresholds to reduce the claim costs (e.g. $5K, $25K, $50K)

Return to Work Model identifies claimants with a long return to work period to assist them to return to work earlier (e.g. 60 days, 90days, 120 days)

Medical Escalation Model predicts claims that have a small medical incurred loss in the beginning but get escalated into large medical incurred loss. (e.g. Difference in incurred loss between 30 days and 2 years >=$2.5K)

Loss Reserve Model predicts the loss reserve amounts for claims when they are closed so that adequate reserve amounts could be booked properly and timely

Litigation Propensity Model Identify a set of claims where the outcomes could be improved by either reducing the likelihood and impact of litigation or a high contentious situation (that might not rise to the level of litigation) but still negatively impact outcomes

Fraud Detection Model predicts and detects the possibility of presence of fraud in order to prevent fraudulent claims

Recovery Model identifies claims with potential for successful subrogation, salvage and refund recoveries exceeding a threshold to improve recovery yield

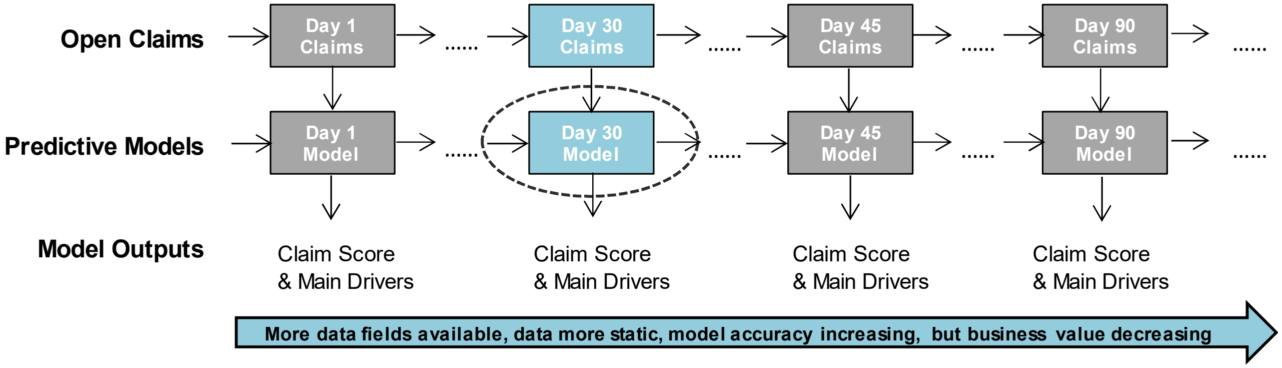
Clinical Guidance Model predicts if the claims need a nurse care management referral to guide a proper treatment

Supervisor Focus Model assists supervisors to identify those claims that look small in the early stage but develop to much larger claims later on

- Insurance Claims Analytics

Near Real-Time Claim Predictive Analytics

Given the claim handling process best practice and associated data collection, building a series of models (corresponding with the associated claim process time lines) to score real time open claims can improve model performance at different time lines which can optimize cost and benefit for the claim handling unit.



Open Claims
Predictive Models
Model Outputs
Day 1 Claims
Day 30 Claims
Day 45 Claims
Day 90 Claims
Day 1 Model
Day 30 Model
Day 45 Model
Day 90 Model
......
Claim Score & Main Drivers
Claim Score & Main Drivers
Claim Score & Main Drivers
Claim Score & Main Drivers
More data fields available, data more static, model accuracy increasing, but business value decreasing

Confidential and Proprietary

Recording of this session via any media type is strictly prohibited

Advanced Analytics Consulting Services, LLC
(AA Consulting)

AGENDA

- ➢ Introduction to Advanced Analytics and Machine Learning
- ➢ Insurance Claims Analytics
- ➢ Return-to-Work Day 30 Model
- ➢ Q & A

Executive Summary

➢ Return-to-Work Day 30 Model

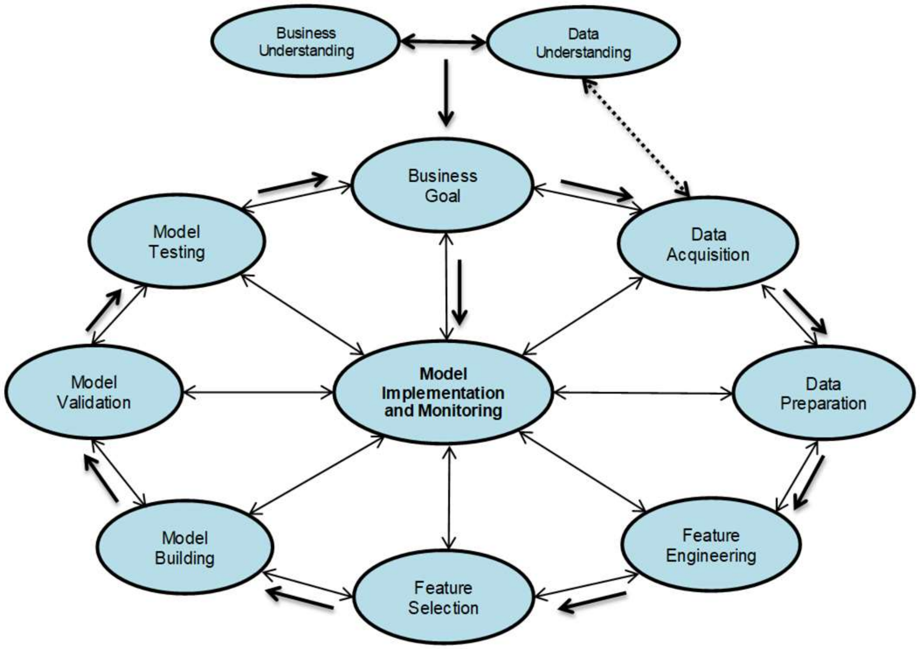
- Motivation - Based on the insurance industry data, the prolonged return to work is one of the main drivers of increased duration and total cost, the sooner the injured worker return to work, the lesser suffering to injured workers and the lower the total claim cost

- Objective - Identify a set of claims where the return to work would be prolonged after day 30 since claim being opened and the outcomes could be improved by more efficient claim handling process, proper treatment, and assistance to return to work.

- Benefit - Improved claim outcomes measured using impact on cost and duration (need to index/severity adjust), Claimant satisfaction.

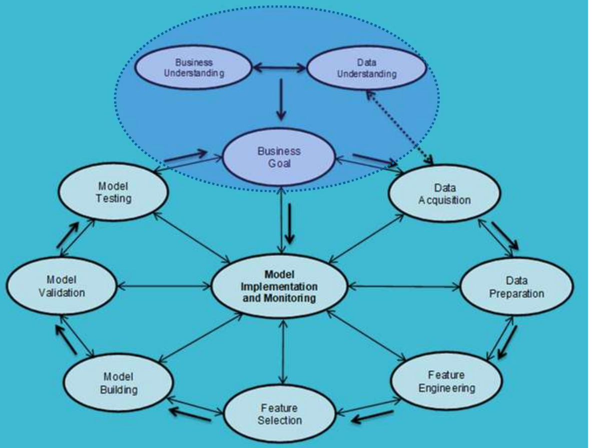
Advanced Analytics Consulting Services, LLC
(AA Consulting)

A Life Cycle Predictive Modeling Process Overview



Note: A standard Property & Casualty insurance predictive modeling process flow chart

- Return-to-Work Day 30 Model



Step 1. Business Goal(s) and Model Design

Objectives:

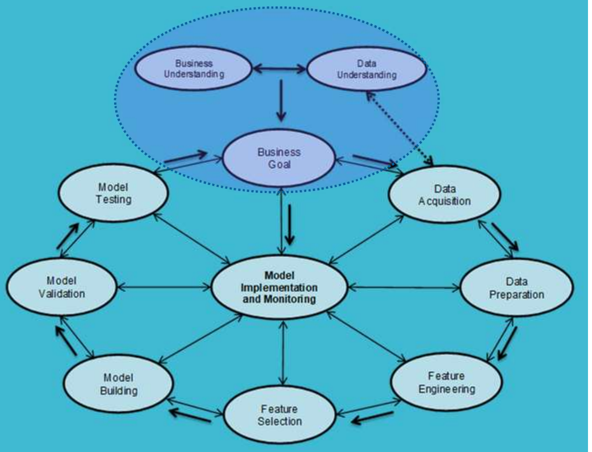
- The business goal is to identify open claims with a high probability of return to work after day 30 since claim being opened in order for claim adjusters to help injured workers return to work earlier
- Model design is to build a return to work model with a binary target variable (Yes/No) to predict the likelihood of injured worker return to work as of day 30.
  - Target variable creation

Challenge:

- A few return to work dates including partial return and full duty return, etc.
- A return to work after day 30 flag proxy could be created depending on what really matters to the company’s business goals

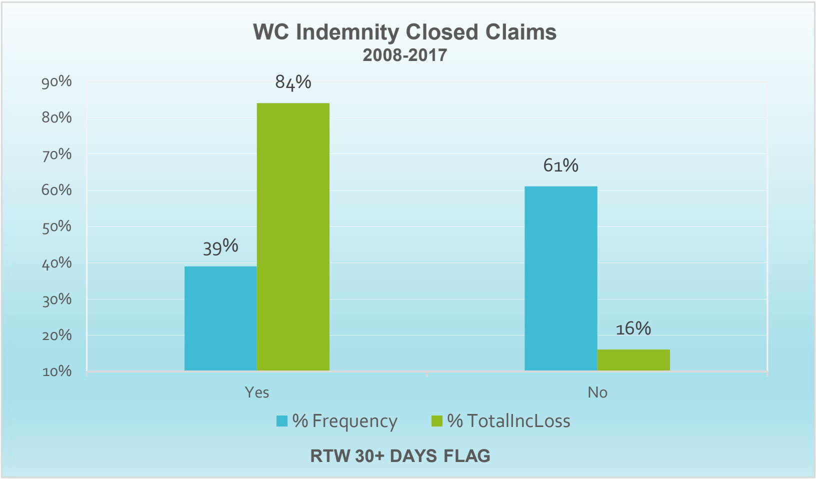
Step 1. Business Goal(s) and Model Design

- Return-to-Work Day 30 Model

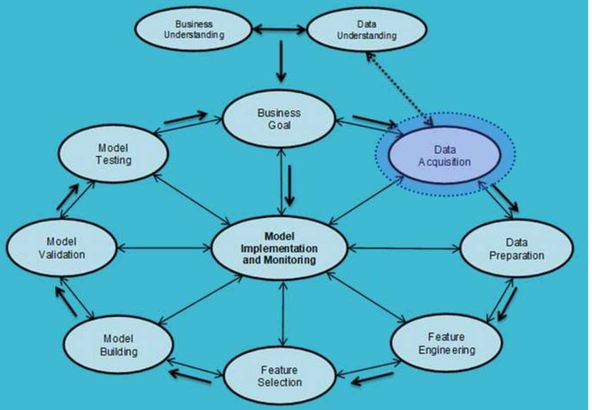


Binary Model and Target Variable Creation:

- Use RTW 30+ days flag directly based on existing solid RTW date data field; Or, create a RTW 30+ days flag proxy based on a combination of RTW date related data fields
- Example: 39% Frequency represents 84% Total Incurred Loss



- Return-to-Work Day 30 Model



Step 2. Data Scope and Acquisition

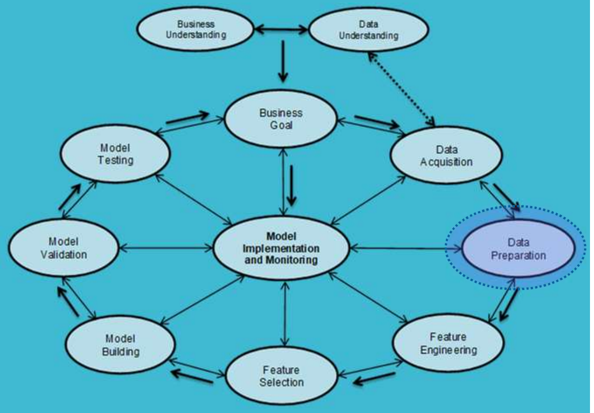
<u>Objectives:</u>

- Data scope:
  - Coverage code = “WC”
  - 10 years WC indemnity closed claims
  - Client status =“current”, etc.
- Data acquisition:
  - Accident, claim, claimant, payment, managed care, demographic, etc.

<u>Rule of thumb:</u>

- If the rare claims/outliers are possible to randomly happen again, then don’t simply exclude them as claims with high severity would impact/drive the overall average loss cost/pricing in insurance data;
- We would like to see not only median but also mean in insurance business for the same reason while some academic research might focus on median for certain research topics.

➢ Return-to-Work Day 30 Model



Step 3. Data Preparation

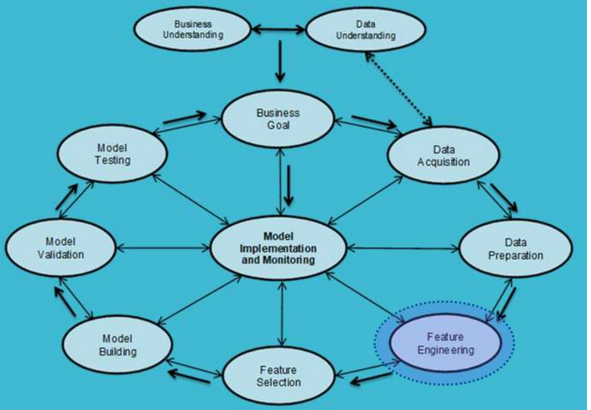
Objective:

- Based on the business goal(s) and data scope, data was reviewed, cleansed, imputed, transformed to be prepared for the next step - variable creation
  - Univariate analysis/Descriptive analytics/Diagnosis analytics
  - Conduct year-to-year, by state, etc. trend study to apply to financial data fields

Examples:

- Cleansing: Total incurred > $0, exclude terminated clients, etc.
- Imputation: Address the missing values (Age, Bill Audit, etc.)
- Transformation: Taking a log or square root or exponential, etc. if data is skewed

- Return-to-Work Day 30 Model



Step 4. Feature Engineering (a.k.a.: Variable Creation)

Objective:

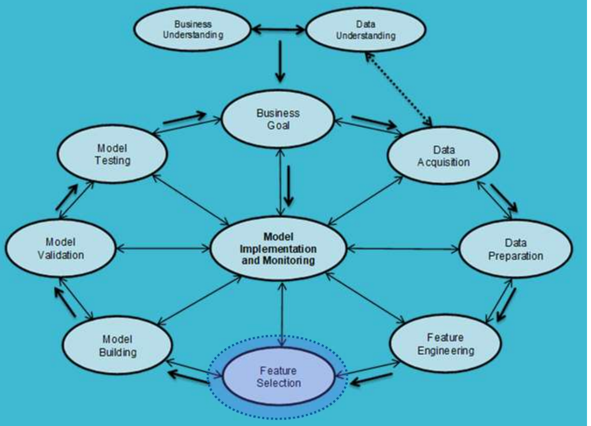
- Create variables that make both statistical and business sense

Examples:

- Data fields could be used directly
  - Initial treatment, Number of dependents, Gender, Marital, Age, etc.
- Create new variables
  - Lags: Based on dates
    - Lags between dates =(max medical improvement date - accident date)
    - Month and Week of accident date, etc. to capture seasonality
  - Groups:
    - Body part group/Injury type group
    - Comorbidity group based on ICD and CPT codes
- Text Analytics to create “variables” based on unstructured data
  - Text Analytics uses algorithms to derive patterns and trends from unstructured (free-form text) data through statistical and machine learning methods as well as natural language processing techniques

Step 5. Feature Selection (a.k.a.: Variable Selection)

- Return-to-Work Day 30 Model

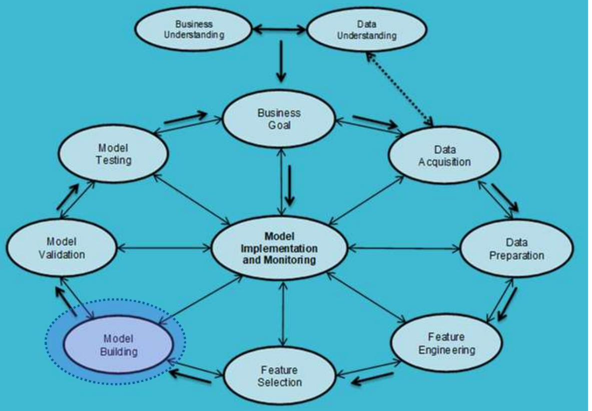


Objective: In machine learning as the dimensionality of the data rises, the amount of data required to provide a reliable analysis grows exponentially so feature learning and selection become critical.

Example:

- To reduce 500+ variables to a manageable level before applying the machine learning algorithms
  - Variable profiling/screening: Missing value ratios, etc.
  - High correlation filters: Identify variables which are correlated to each other to avoid multicollinearity
  - Multivariate analyses: cluster analysis, principle component analysis, and factor analysis, etc.
  - Stepwise
- There are also some built-in variable selection methods depending on specific type of statistical tools (SAS, R, and Python)

- ➢ Return-to-Work Day 30 Model



Step 6. Model Building (a.k.a.: Model Fitting)

Objective:

- After serious data mining work, more than one statistical software used along with multiple machine learning algorithms utilized to build model to have a few candidate models (usually 3)
- Interaction and correlation usually should be examined before finalizing the models

Example of Model Comparison (accuracy, precision, and implementation considerations):

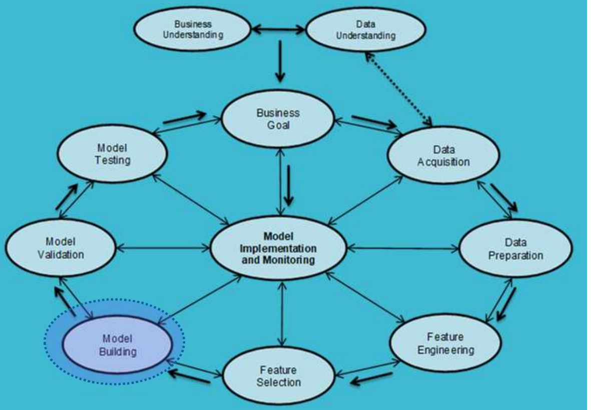
Measurement	GLM Logistics Regression	Decision Tree	Random Forests	Gradient Boosting	Neural Network	Support Vector Machine
Accuracy Ratio	89%	89%	89%	87%	90%	89%
Precision Ratio	74%	74%	74%	73%	75%	74%

Example of Main Drivers:

- Lag of max medical improvement, Injury type, Certain type of body part code, Comorbidity group, Nature result group, Average weekly wage, Benefit state, Higher than average number of treatments, More than 1 medical provider, Union flag, Industry, etc.

Step 6. Model Building – Data Partition

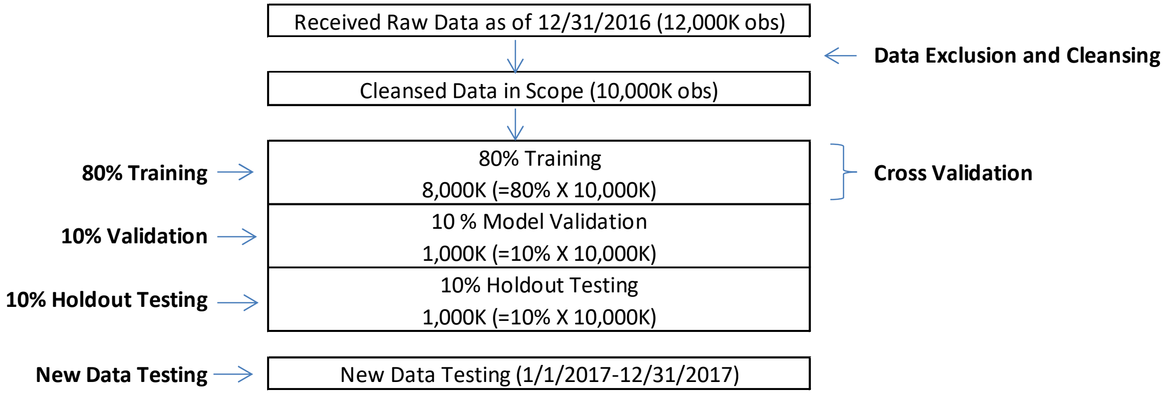
- Return-to-Work Day 30 Model



Objective:

- Ideally, the data would be partitioned into training, validation, and testing data sets.

Example of Data Partition:



Received Raw Data as of 12/31/2016 (12,000K obs)

Data Exclusion and Cleansing

Cleansed Data in Scope (10,000K obs)

80% Training → 80% Training 8,000K (=80% X 10,000K)

Cross Validation

10% Validation → 10 % Model Validation 1,000K (=10% X 10,000K)

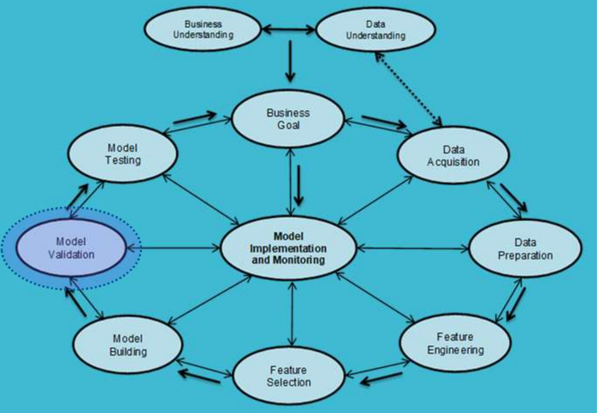
10% Holdout Testing → 10% Holdout Testing 1,000K (=10% X 10,000K)

New Data Testing → New Data Testing (1/1/2017-12/31/2017)

Notes: Training data: Data used to build model
Validation data: Data used to validate and select the best model from model candidates
Testing data: Data used to test model before implementation

Step 7. Model Validation

- Return-to-Work Day 30 Model



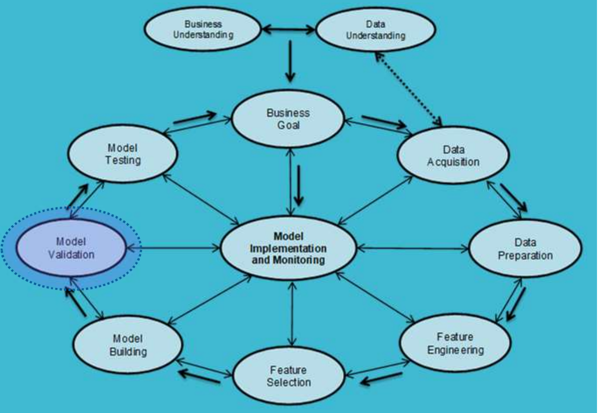
Objective:

- Model validation is a process to apply the model on the validation data set to select a best model from candidate models with a good balance of accuracy and stability

- Common validation methods includes cross validation, lift charts, confusion matrices, receiver operating characteristic (ROC), and bootstrap sampling, etc. to compare actual values (results) versus predicted values from the model.
  - Bootstrap sampling and cross validation are especially useful when data volume is low

Step 7. Model Validation

- Return-to-Work Day 30 Model

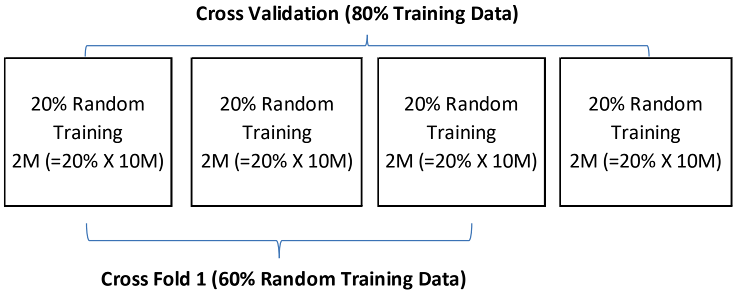


Cross-validation:

- Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

  In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

- Simple Example of 4-fold Cross-validation:



Cross Validation (80% Training Data)

20% Random Training 2M (=20% X 10M)

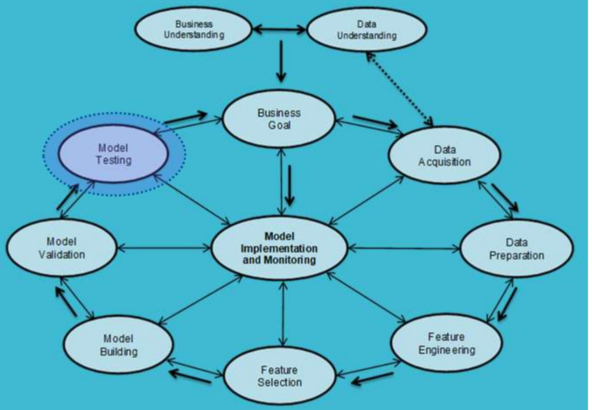
20% Random Training 2M (=20% X 10M)

20% Random Training 2M (=20% X 10M)

20% Random Training 2M (=20% X 10M)

Cross Fold 1 (60% Random Training Data)

- ➢ Return-to-Work Day 30 Model



Step 8. Model Testing

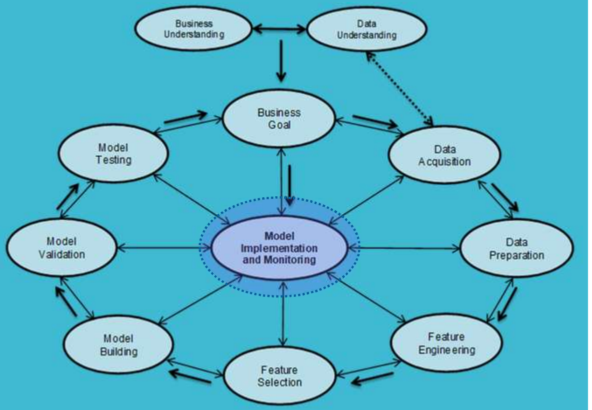
Objective:

- Model testing is performed by using the best model from the model validation process to further evaluate the model performance and provide a final honest model assessment
- Model testing technical methods are similar as model validation but using holdout testing data and/or new data

Examples:

- Test on hold-out data sets
- Test on newly open claims
- Test by major clients
- Test by major industries
- Test by benefit states
- Test by major injury types

- ➢ Return-to-Work Day 30 Model

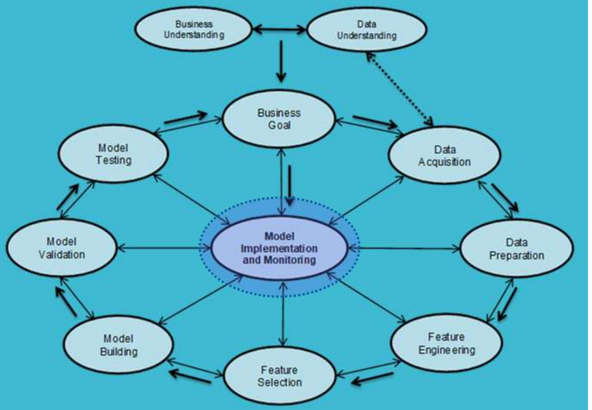


Step 9. Model Implementation

Objective:

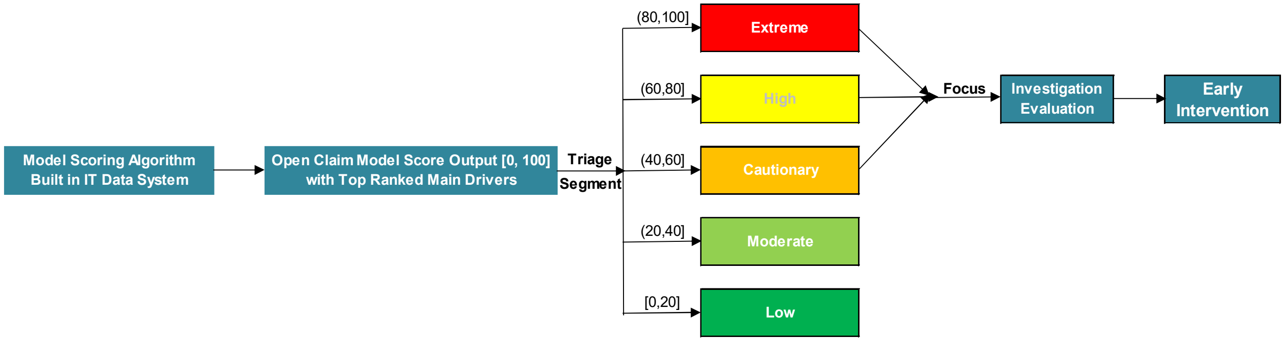
- ▪ The last but important step is the model implementation:
  - Turn all the modeling work into action to achieve the business goal and/or solve the business problem
  - Before model implementation, a model pilot would be helpful to further confirm the model performance and prepare for implementation appropriately
  - If there is an existing model, we would like to conduct a model champion challenge to understand the benefit of implementing the new model over the old one
- ▪ Model performance monitoring and results evaluation should be conducted periodically

➢ Return-to-Work Day 30 Model



Step 9. Model Implementation

- An example of technical implementation



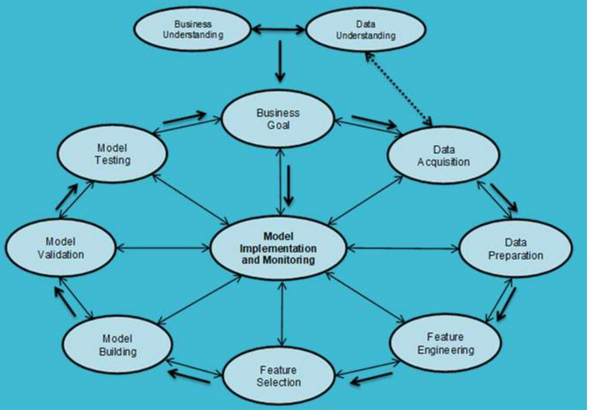
A data-driven opinion to assist claim professionals:

1. Gain insights to the claims based on claims' model score and the main drivers
2. Stay focused on the claims with higher scores/high propensity to improve claim handling efficiency
3. Optimize claim handling resources by assigning claims to adjusters with proper experience
4. Close the claims earlier with lower cost to drive demonstrably superior claim outcomes

Note:
After implementing the model, the model need to be reviewed and rebuilt periodically, such as annually.

Lessons Learned

- Return-to-Work Day 30 Model



- **A successful model is team work across functional departments –** Advanced analytics consultant/predictive modeler/data scientist, IT support, BI, data management, and business partners' inputs together are the key to the model success. Statistical models, machine learning algorithms, data knowledge, and business domain knowledge should always be tied to the business goals to build an implementable models.

- **Rome wasn't built in a day, and the same is true with a superior predictive model and advanced analytics capability** – Start to build an initial model with proof of concepts and gradually improve the process and the model. Develop one model first, then add more models as of different ages to improve the results over the time.

- **Do not boil the ocean -** Avoid the temptation to use all of the available data to build the initial version of the model. Generally, start with readily available structured data (e.g. claim, claimant, accident, and payment), then extend to unstructured data, semi-unstructured data and add more data channels (e.g. external data sources) to future iterations of the models. Avoid trying all the available new modeling techniques initially.

- **Garbage in garbage out -** Historical relevant data is far from perfect, we need to take time to understand the data (quality and relevance), process it, and get it ready for machine learning. Over 80% of the total time is spent on preparing the data for the analysis.

- **Interpretability is the key to success -** Model users are not the same as the model builders. Keep the outputs simple, business meaningful, and usable for your users.

Advanced Analytics Consulting Services, LLC
(AA Consulting)

Questions & Answers

Thank You!

Contact Information:

Your comments and questions are valued and encouraged.

Name: Mei Najim, CSPA, Founder and Advanced Analytics Consultant
Company: Advanced Analytics Consulting Services, LLC

E-mail: mei_najim@aacsus.com

LinkedIn: https://www.linkedin.com/in/meinajim/

Website: https://www.aacsus.com/