Applying Predictive Modeling to Auto Insurance Pricing Optimization

CAS Spring Meeting
May 2002

Moderator: **Lynne Bloom**

Presenters: **Peter Phillips**, President and CEO of Aon PathWise
Kailan(Kevin) Shang, Associate Director of Aon PathWise

Prepared by PathWise® Solutions Group

Empower Results®

Legal Disclaimer

This presentations, data, information and all other contents or materials contained herein or attached hereto (collectively, the "Materials") were intended and prepared for general informational purposes only and should not be construed as advice or opinions on any specific facts or circumstances. These Materials are made available on an "as is" basis without any warranty of any kind (express or implied), including without limitation respect to the accuracy, completeness, timeliness, or sufficiency of these Materials. PathWise Solutions Group LLC (“PSG LLC”) and its affiliates, directors, officers, employees, and representatives (collectively, “Representatives”) disclaim any liability to any person or organization for loss or damage caused by or resulting from any actions based on or reliance placed on these Materials. Neither PSG LLC nor its Representatives will be liable, in any event, for any special, indirect, consequential, or punitive loss or damage of any kind arising from or relating to these Materials. These Materials are intended only for the designated recipient to whom they were originally delivered and any person or organization in receipt of, or otherwise accessing, these Materials shall not provide or make available these Materials, or any portion or summary hereof, to any third party without the express written consent of PSG LLC. PSG LLC does not provide and these Materials do not constitute any form of legal, accounting, taxation, regulatory, or actuarial advice. The recipient of these Materials is advised to undertake an independent review of any legal, accounting, taxation, regulatory, and actuarial implications of any transaction described herein.

Agenda

Agenda

What is Predictive Modeling?

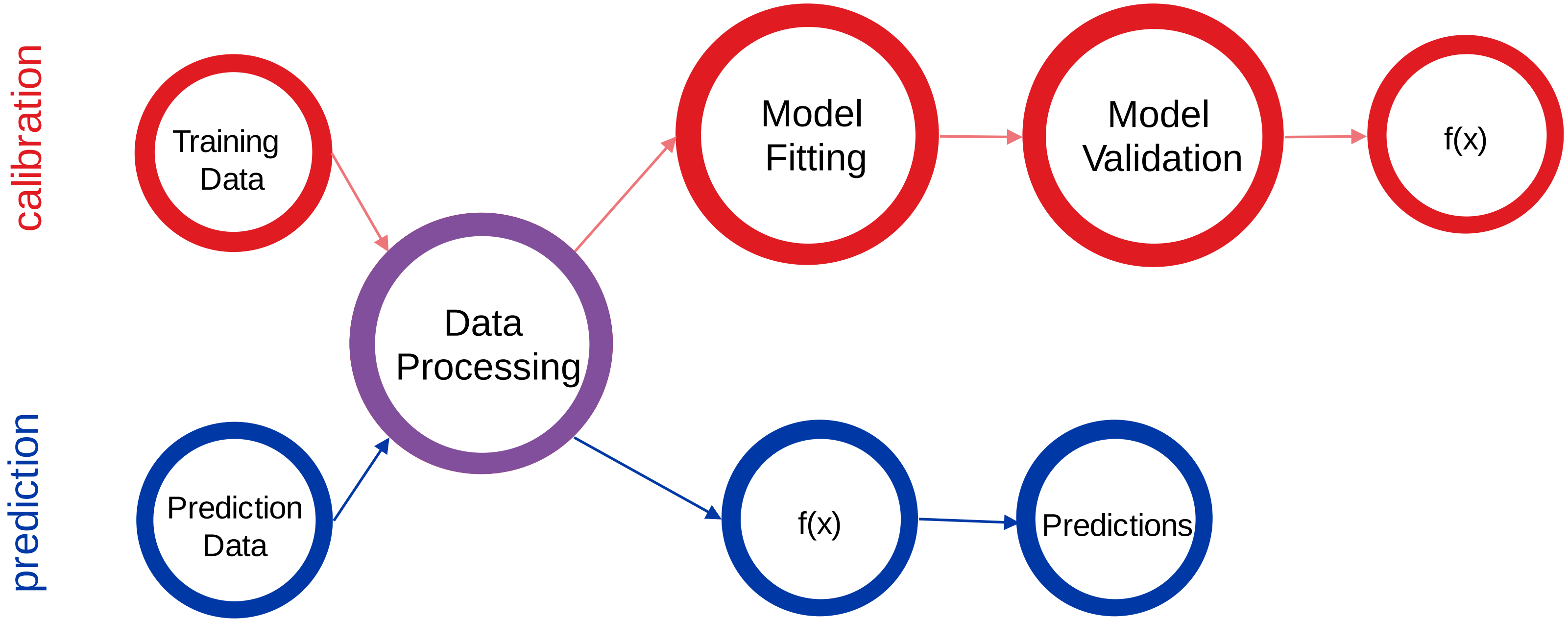
Predictive modeling is a commonly used **statistical technique** to **predict future behavior** by **analyzing** historical and current **data** and generating a model.

Driving Forces	Data volume Computing capabilities
Model Type	Linear regression/GLM → RandomForest/GBM/Artificial Neural Networks
Algorithm	$Y = X\beta + \varepsilon$ $\beta = (X'X)^{-1}X'Y$ Inverse matrix operation → gradient descend method
Validation	Hypothesis tests → prediction accuracy using out-of-sample data

Applications

Pricing	Reserving	Claim	Risk Management
• Rate setting at policy level • Driving behavior Analysis • Underwriting decision-making	• Case reserving • IBNER development pattern prediction • Salvage • Subrogation	• Open claim classification • Claim decision-making for small cases	• Fraud detection

Predictive Modeling Process



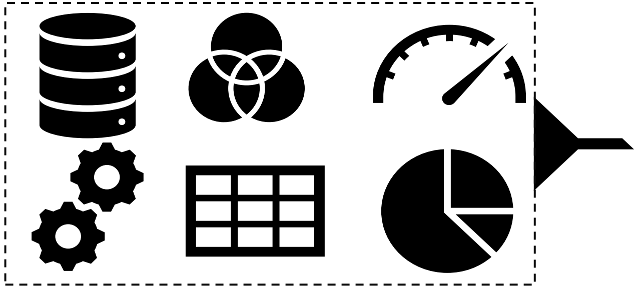
Training Data

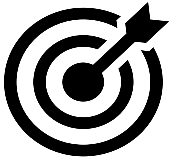
Policy Info	Demographic info, financial info, insured property, deductible, limit, ...
Claim Info	Date, time, location, severity, reporting lag, settlement lag, adjuster's assessment, ...
LAE	Loss adjustment expense
Market	Soft vs. hard market, inflation, ...

Data Preparation

Data is more important than models nowadays. Everyone can run models.

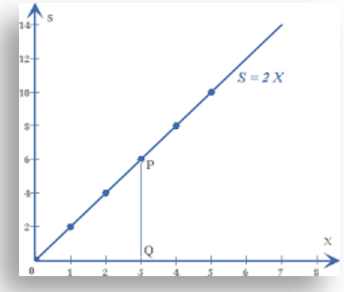
Variables	Data Validation	Feature Engineering	Dimensionality Reduction
• Convert categorical variables to dummy variables • Text Mining	• Missing data treatment • Scaling • Constant variable	• Create new variables to reflect nonlinear relationships	• Principal component analysis • Collinearity



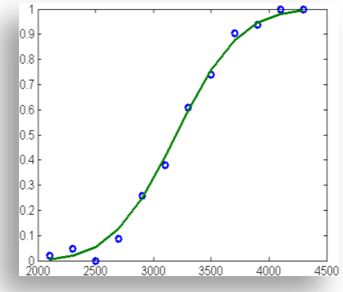


Model Choices – Supervised and Unsupervised Learning

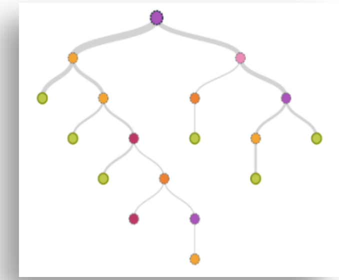
Linear Regression



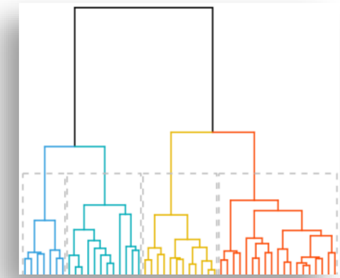
GLM



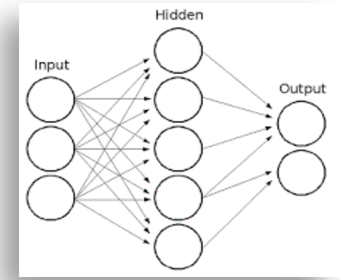
CART



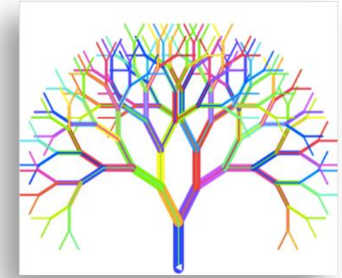
Hierarchical Clustering



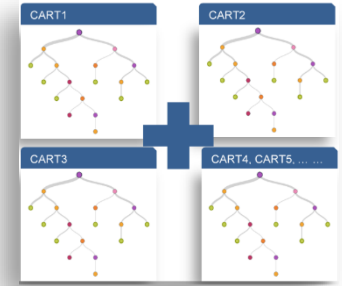
Neural Networks



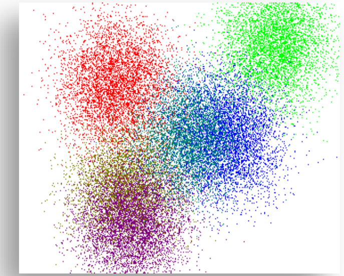
RandomForest



GBM



K-Means



And the list goes on and on … …

Model Fitting

Error Function	Overfitting	Hyperparameter	Validation
• RMSE • MAE • Weighted RMSE • Huber Loss • Quantile Loss	• Regularization (Lasso, Ridge, and Elastic Net) • Random data subset • Random feature subset • Neuron dropout	• Size of random subset • Learning rate • Depth of tree models	• Training/validation split • Cross Validation

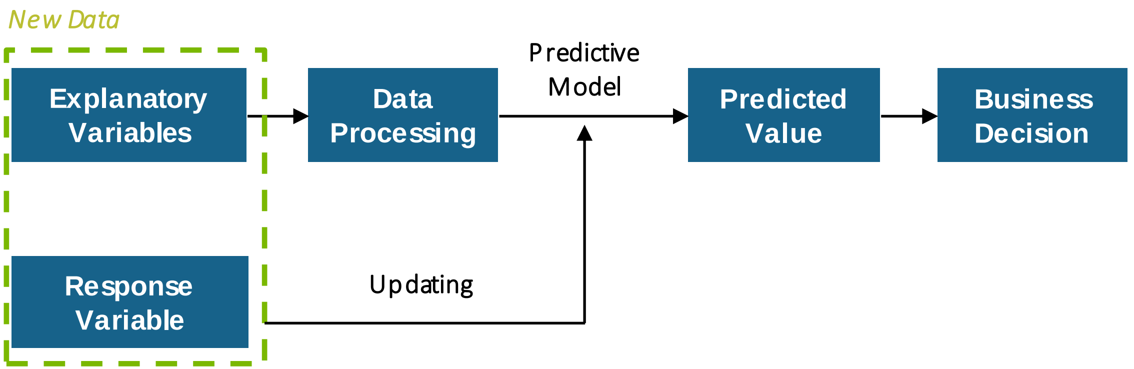
Model Validation

Out-of-sample data is used for model validation

	Considerations
Goodness of Fit	• R-squared • Adjusted R-squared • Precision, Recall, F-measure • AUC (Area under the curve) Regression Classification
Outliers	• Scatter plot • Predictions with error outside (m-3s, m+3s)
Feature Importance	• Most important variables

Model validation is the key to building knowledge and confidence in complex models

Program Maintenance



- If the new data exhibits similar distributions and relationships to the existing data, model updating is not necessary.
- A threshold of new data volume may be set to trigger the updating process.
- Exclude variables whose volatility has been reflected in the training data
- Consistency with the usage of predictive modeling
- Automation is the key to efficient implementation

Challenges

Knowledge gap for complex models

Translate higher accuracy into better strategies

Multiple model types and fine tuning

Get stakeholder buy-in

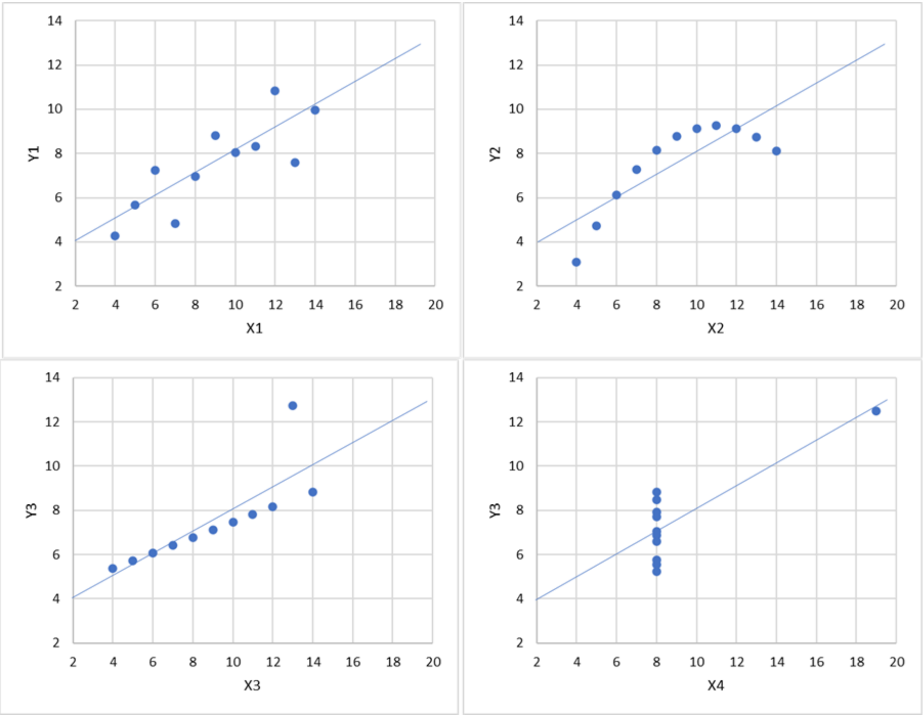
Parallelization of model training

Mature validation process

Automation of training and validation

Example: Anscombe Quartet

$X_{1,2,3}$	Y_1	Y_2	Y_3	X_4	Y_4
10	8.04	9.14	7.46	8	6.58
8	6.95	8.14	6.77	8	5.76
13	7.58	8.74	12.74	8	7.71
9	8.81	8.77	7.11	8	8.84
11	8.33	9.26	7.81	8	8.47
14	9.96	8.1	8.84	8	7.04
6	7.24	6.13	6.08	8	5.25
4	4.26	3.1	5.39	19	12.5
12	10.84	9.13	8.15	8	5.56
7	4.82	7.26	6.42	8	7.91
5	5.68	4.74	5.73	8	6.89
Mean	7.50	7.50	7.50		7.50
Standard Deviation	2.03	2.03	2.03		2.03
Correlation with X	0.816	0.816	0.816	Correlation with X4	0.816
Linear Regression	Y=3+0.5X				
R2	0.666	0.666	0.666		0.666



Empower Results®

Agenda

Section 1	Introduction to Predictive Modeling
Section 2	**Applying Predictive Modeling to Auto Insurance Pricing**
Section 3	Pricing Optimization with Better Prediction
Section 4	Recap and Q&A

Laboratory Setting

Synthetic auto claim data

policy_ID	100-001	100-002	100-003	100-004	100-005
policy_age	0	23	1	7	11
num_drivers	1	2	1	1	1
mileage	98431	99166	4403	70952	201235
primary_driver_age	45	80	23	36	32
primary_driver_gender	male	female	female	male	male
occupation_ID	occ_#1	retired	occ_#2	occ_#2	occ_#4
region	city_#8	city_#2	city_#6	city_#9	city_#7
vehicle_type	veh_type1	veh_type6	veh_type3	veh_type5	veh_type2
vehicle_power	pow_type0	pow_type2	pow_type1	pow_type2	pow_type2
usage	work_private	retired	work_private	work_private	commercial
no_of_past_claims	1	no	no	no	no
past_severity	327.4	0	0	0	0
is_loss (target - frequency)	**No**	**No**	**Yes**	**No**	**No**
loss_amount (target - severity)	**0**	**0**	**4,530**	**0**	**0**

Data Processing

	Frequency	Severity
Data Record	100,000	2,572 (Frequency = 1)
Data Type	0 or 1	Loss with a limit of 100,000
Model	Classification	Regression
Descriptive Statistics	Avg.: 2.572%	Avg.: 25,942 Std: 12,655

Auto Premium = Frequency x Severity

- Categorical variables are converted to dummy variables
- Missing data is removed
- Mileage and past severity are scaled to range [0,1]
- Correlation analysis is performed to identify highly correlated pairs

Example: Supervised Learning

1. *Predict the probability of have an insurance claim at policy level*

2. *If an insurance claim is predicted, predict the claim amount*

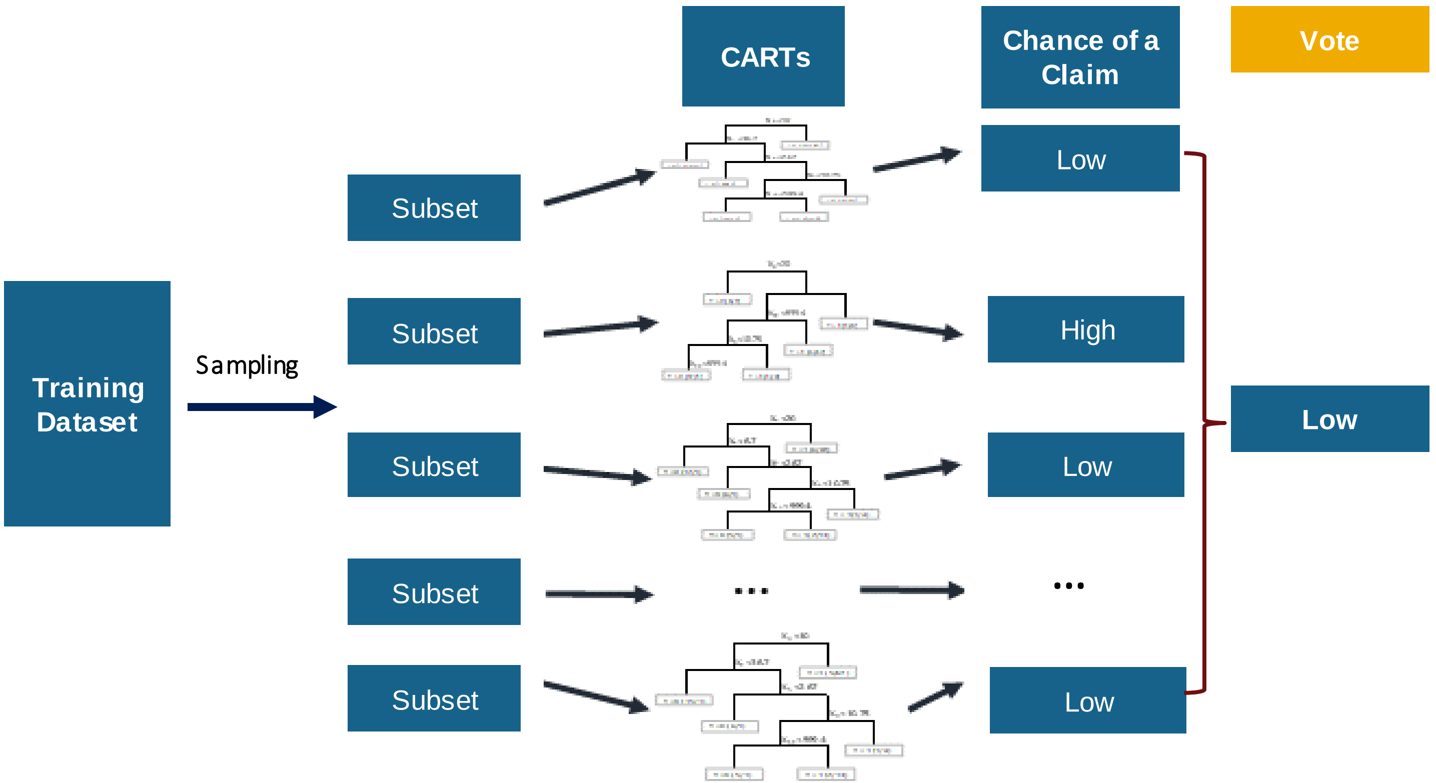
GLM

$$\mathrm{E}(Y|X) = \mu = g^{-1}(\eta) = g^{-1}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)$$

Logistic Model: a special case of GLM

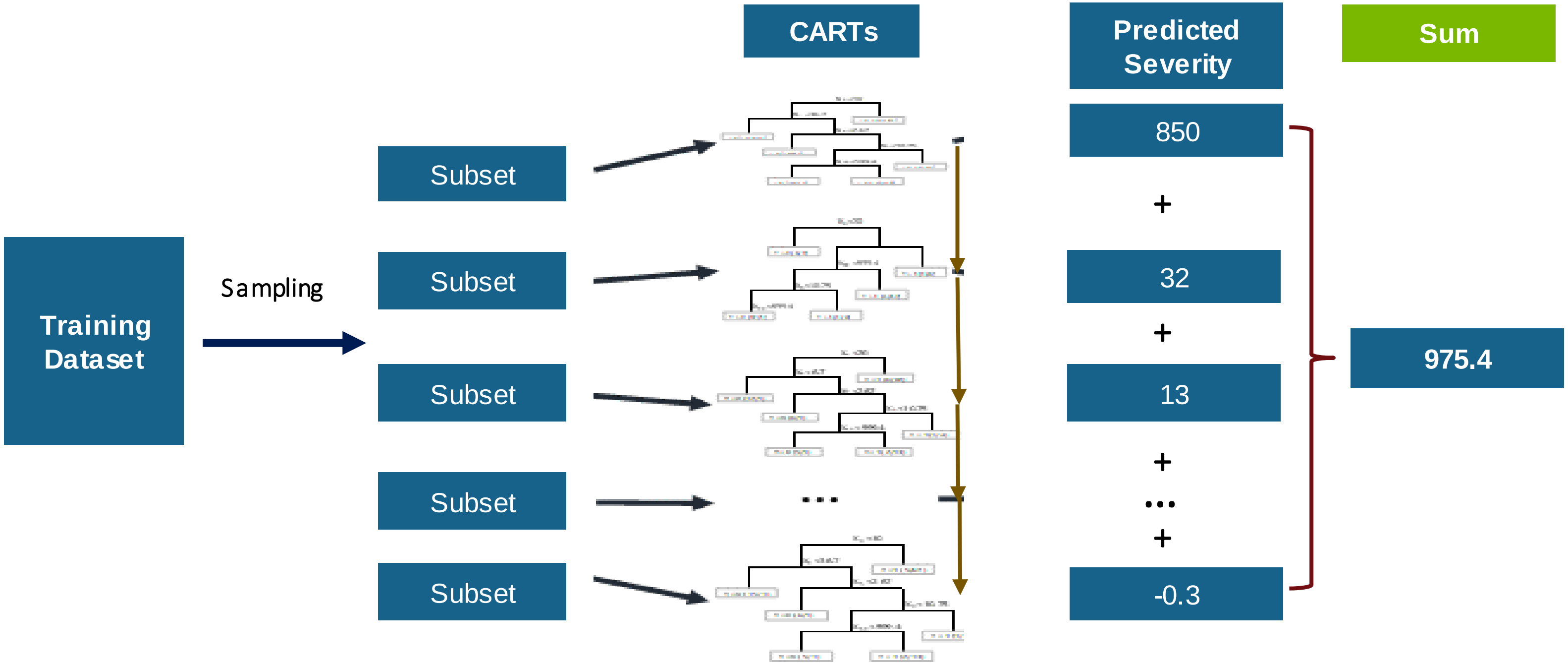
$$\mathrm{E}(Y|X) = \mu = \frac{1}{1 + e^{-\eta}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

Random Forests Model Structure

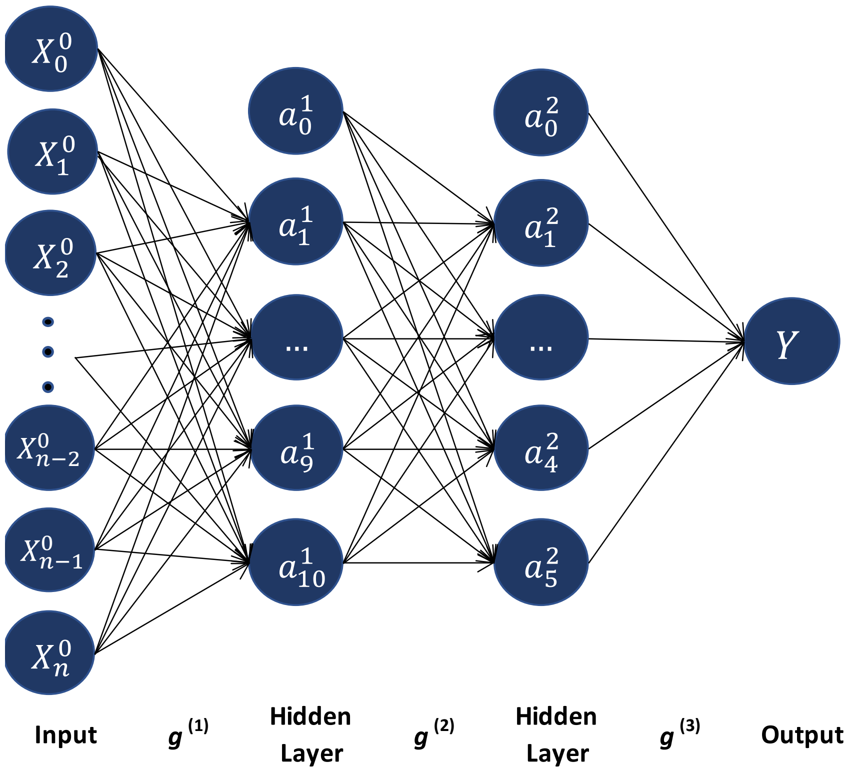


Empower Results®

GBM Model Structure



Artificial Neural Networks Model Structure

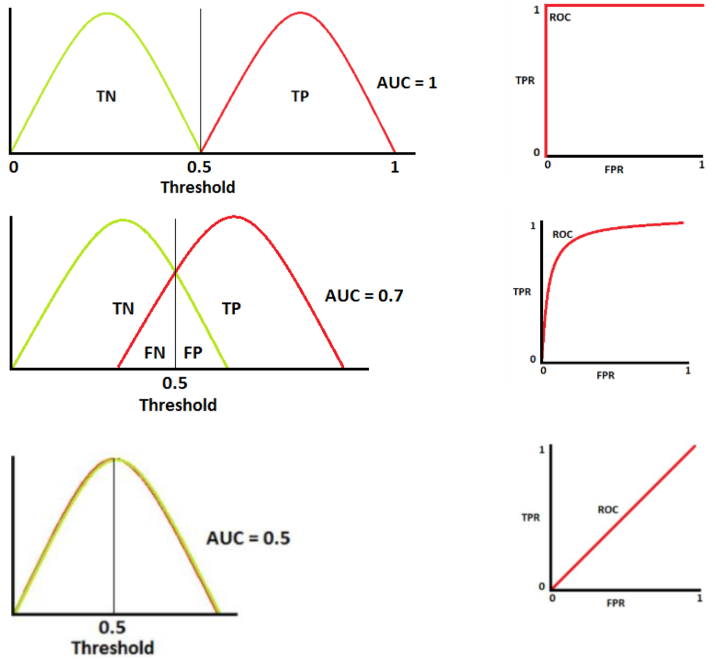


ANN uses multiple layers of linear, logistic or other simple functions to allow many more possible relationships

Calibration Results – Frequency

	AUC
Logistic	81.43%
Random Forest	97.52%
GBM	98.76%
ANN	99.51%

AUC - Area Under The Curve
ROC - Receiver Operating Characteristics curve



- ROC is a probability curve and AUC represents the degree or measure of separability.
- An excellent model with AUC close to 1 has superior measure of separability.
- An AUC of 0.7 indicates there is 70% chance that the model will be able to distinguish between positive class and negative class.

Hyperparameters

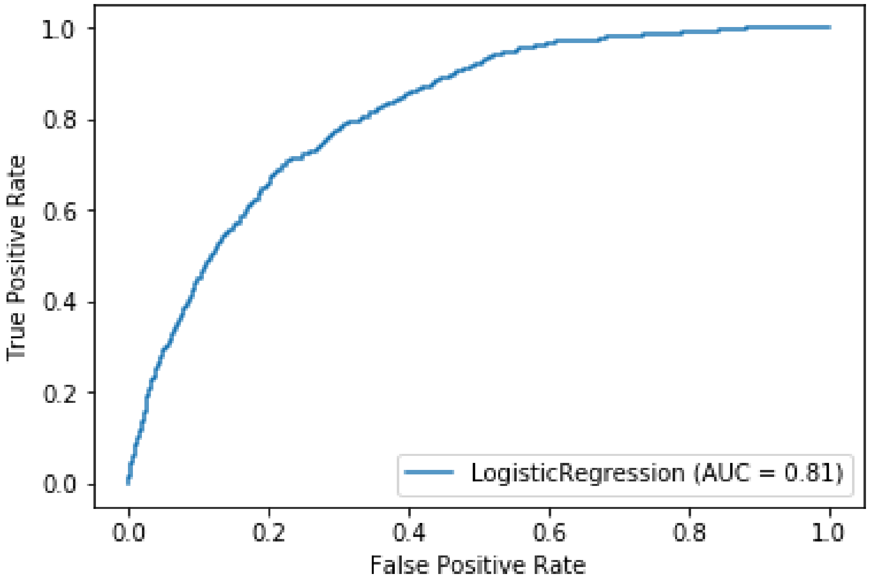
	AUC
Logistic	81.43%
Random Forest	97.52%
GBM	98.76%
ANN	99.51%

Tested GBM Hyperparameters
error function
maximum iteration
batch size
error tolerance
L1 ratio to test for regularized models
L2 ratio to test for regularized models
number of estimators
learning rate
fraction of samples to be used for fitting the individual base learners
max depth of the tree model
minimum number of samples required to split an internal node
minimum number of samples required to be at a leaf node
number of features to consider when looking for the best split

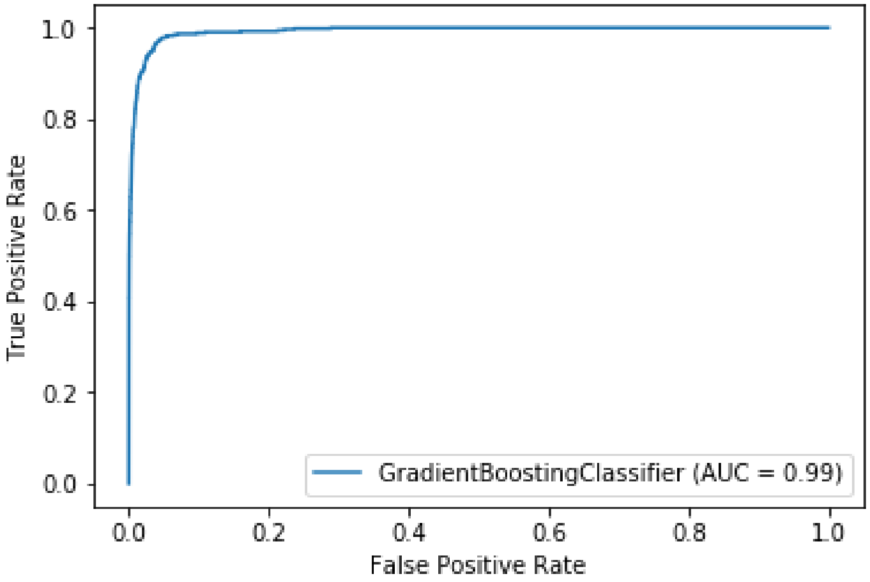
Model Validation – Frequency

AUC – ROC Curve

Logistic



GBM

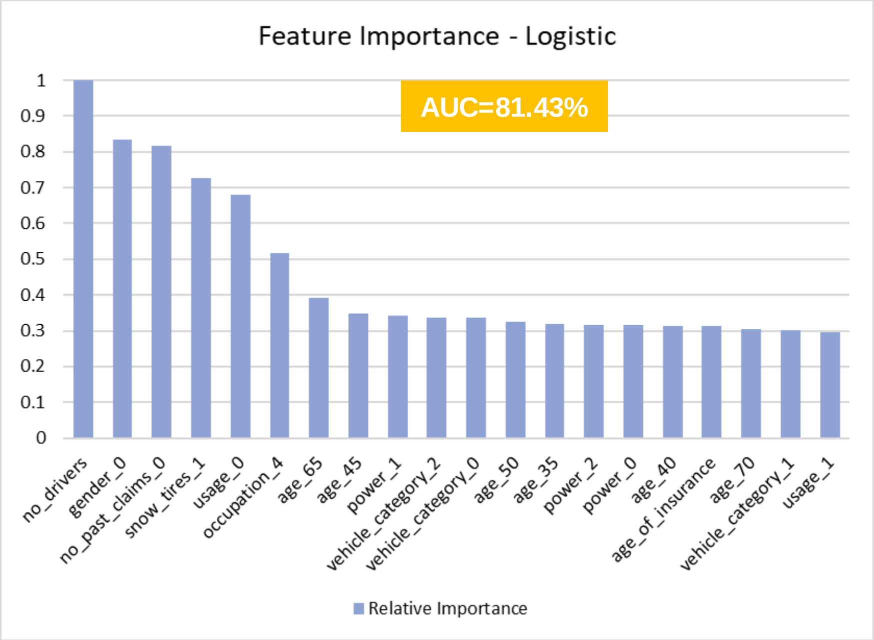


- AUC – ROC Curve indicates the model's capability to distinguish between classes.
- **Validation** data points were not used when calibrating the model.

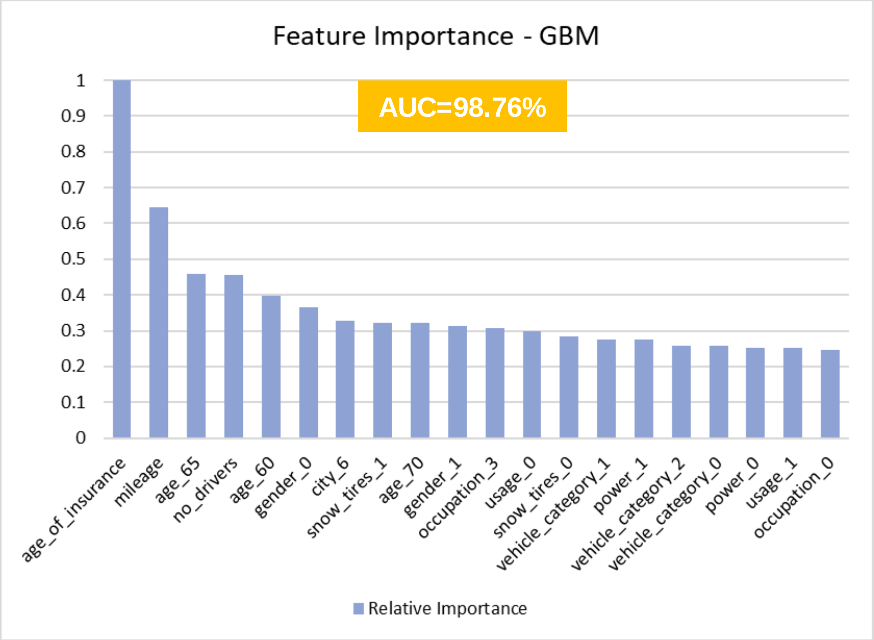
Model Validation – Frequency

Important Features

Logistic



GBM

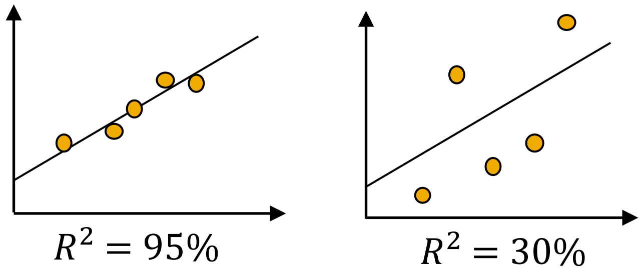


- Important features are similar while the order of importance may change.
- We can identify key risk indicators such as number of drivers and age of insurance.

Calibration Results – Severity

Model	R^2
GLM	90.92%
GBM	93.62%
ANN	93.35%

$$R^2 = 1 - \frac{Unexplained\ Variation}{Total\ Variation}$$

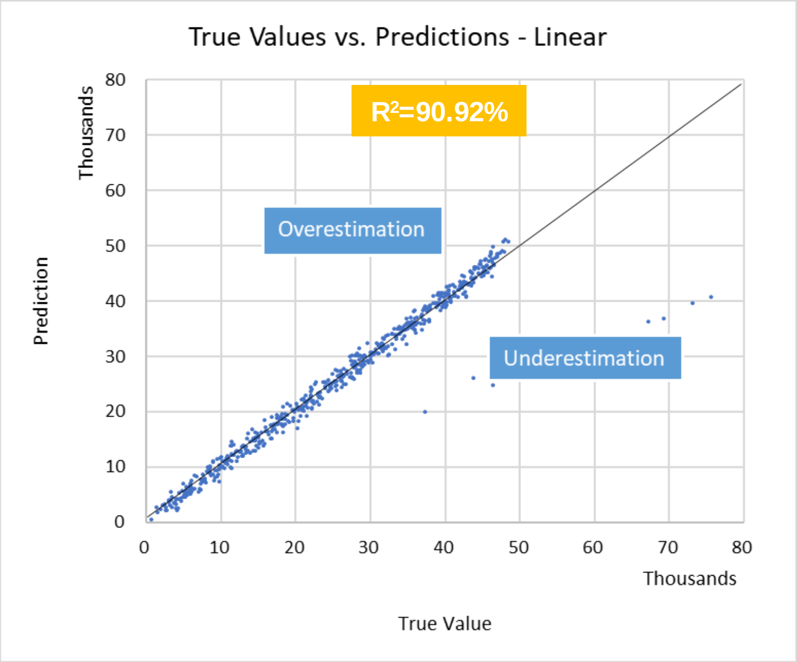


- R^2 is used to measure the prediction accuracy.

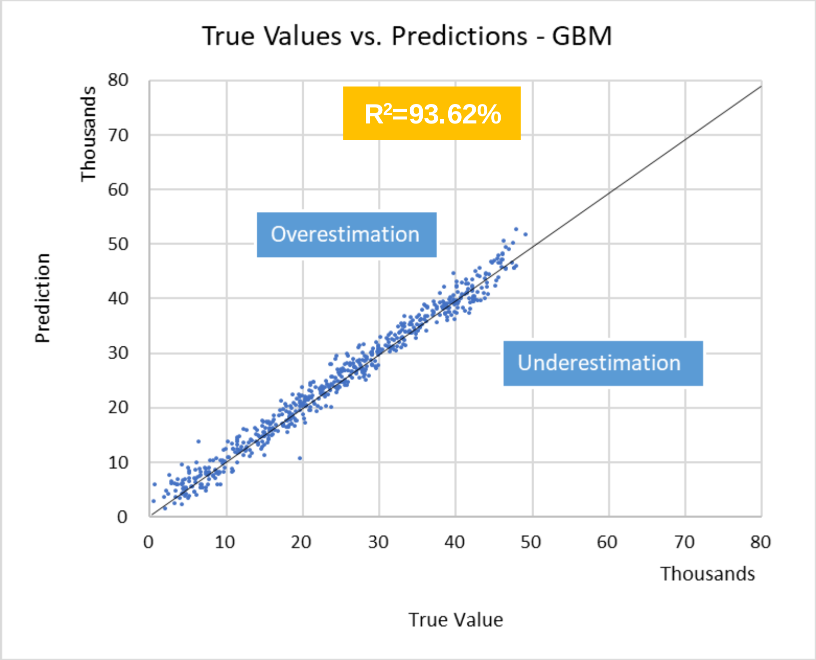
Model Validation – Severity

true values vs. predictions (Out-of-sample data)

Generalized Linear Model



GBM



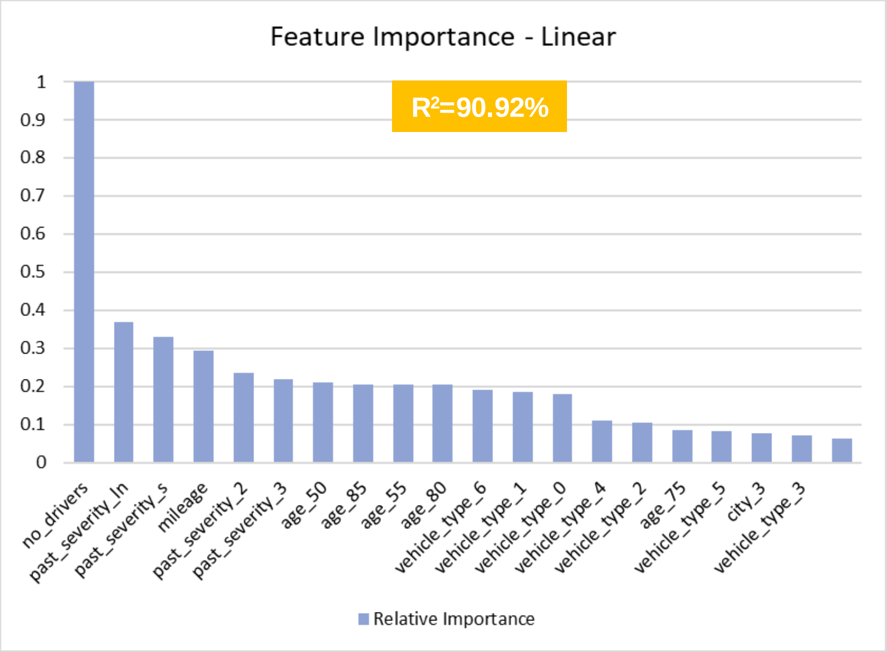
- Scatter plots show Severity values based on true values and prediction
- **Validation** data points varying by outer loop scenario and time point

Empower Results®

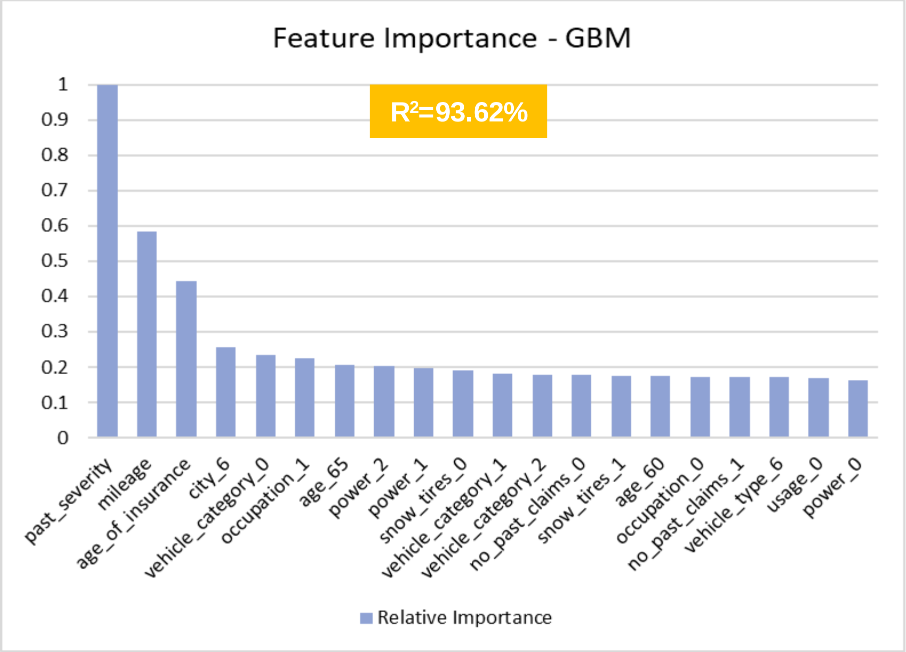
Model Validation – Severity

Important Features

Generalized Linear Model



GBM



- Important features are similar while the order of importance may change.
- We can identify key risk indicators such as past severity and mileage.

Model Selection

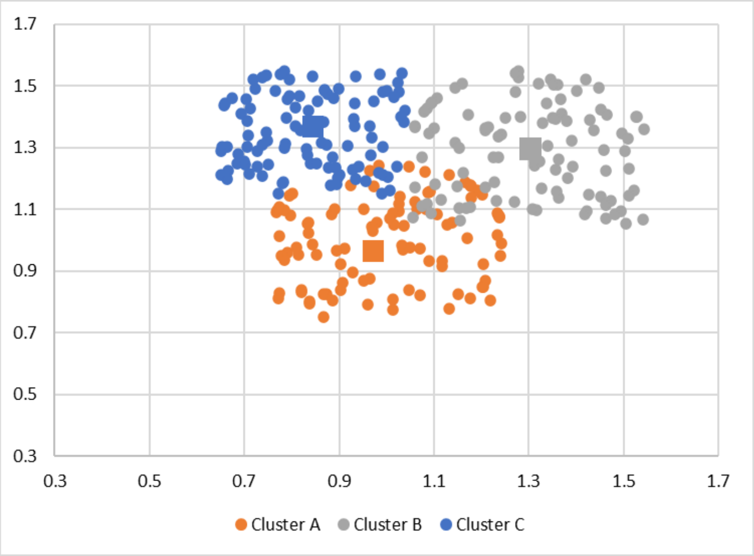
	Linear	Gradient Boosting Machine (GBM)	Neural Network
Features	• Feature engineering to capture non-linear relationships	• An ensemble of weak predictors in the form of decision trees • Each predictor is additive trying to minimize the residual error	• A set of algorithms designed to recognize patterns.
pros	• Easy to understand and validate	• Better prediction accuracy	• Good with nonlinear data with more data points
cons	• When adding new model variables, calibration needs to be refined.	• Exact prediction rule is not very transparent although the accuracy can be backed by validation. • Need to gain knowledge of this model.	• More computationally expensive • More challenging to interpreter the relationships between the independent variables and the dependent variable.

Example: Unsupervised Learning

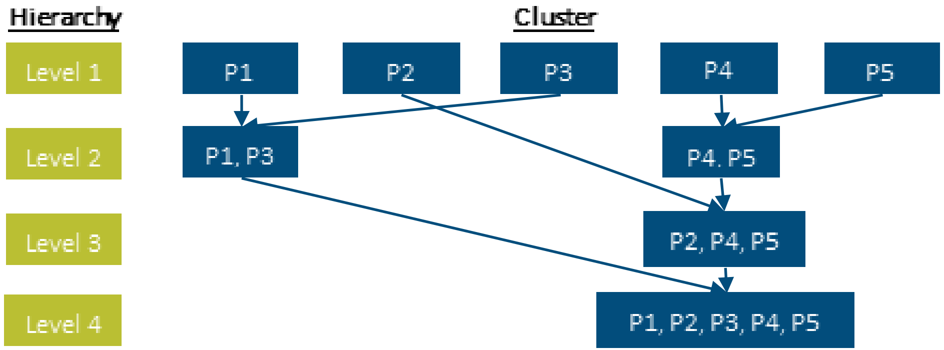
1. *Classify policies based purely on explanatory variables*
2. *Assess the loss probability and risks for each cluster*

Models

K-means



Hierarchical clustering



Unsupervised Clustering Application – Risk Rating

Risk **Cluster**	Loss Probability $\hat{p}$	Number of Members
Only 1	3%	100,000

Risk **Cluster**	Loss Probability $\hat{p}$	Number of Members
#1	0%	52,895
#2	1%	7,692
#3	2%	5,103
#4	3%	4,361
#5	4%	4,461
#6	5%	4,246
#7	6%	4,795
#8	7%	3,334
#9	8%	3,448
#10	9%	2,488
#11	10%	1,930
#12	11%	1,570
#13	12%	1,256
#14	13%	1,022
#15	14%	658
#16	15%	256
#17	16%	356
#18	21%	129

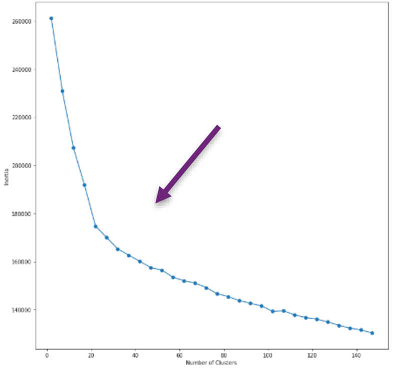
- The above demonstrates an application of the K-Means clustering algorithm on insurance risk rating. This algorithm optimizes the intra-cluster squared errors *(inertia).*
- In this specific example, it divides 100,000 insurance risks into 1,000 clusters / risk cohorts to manage risk at a granular level.

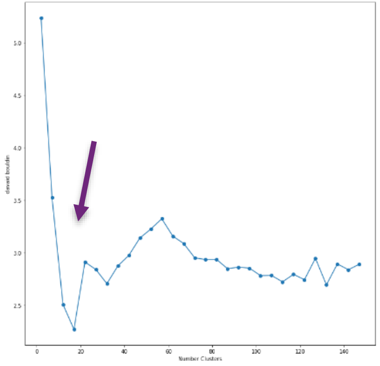
Unsupervised Clustering Application – Risk Rating

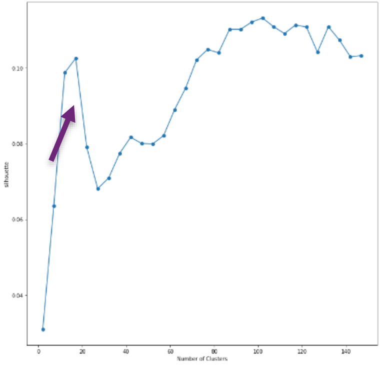
policy_ID	policy_age	num_drivers	mileage	primary_driver_age	primary_driver_gender	occupation_ID	region	vehicle_type	vehicle_power	usage	past_lost	cluster_label	risk_rating	is_loss (target)
100-001	0	1	98,431	45	male	occ_#1	city_#8	veh_type1	pow_type0	work_private	yes	#51	16.45%	No
100-002	23	2	99,166	80	female	retired	city_#2	veh_type6	pow_type2	retired	no	#2	5.15%	No
100-003	1	1	4,403	23	female	occ_#2	city_#6	veh_type3	pow_type1	work_private	no	#16	72.34%	Yes
100-004	7	1	70,952	36	male	occ_#2	city_#9	veh_type5	pow_type2	work_private	no	#8	0.00%	No
100-005	11	1	201,235	32	male	occ_#4	city_#7	veh_type2	pow_type2	commercial	no	#9	10.33%	No

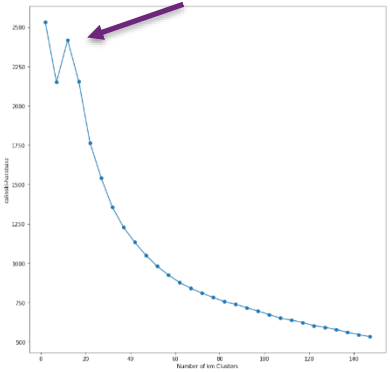
The unsupervised learning algorithm determines the cluster labels, which are then used for risk rating purposes.

Unsupervised Clustering – Determination of the Number of Clusters









Elbow Method

finds the optimal point that balances the model complexity and within-cluster sum-of-squares (*inertia*)

Davies-Bouldin Index

compares the cluster diameters with the distance between cluster centroids for each pair of clusters

Silhouette Coefficient

compares the average intra-cluster distance to the nearest-cluster distance for each point

Calinski Harabasz Score

(a.k.a. the Variance Ratio Criterion)

contrasts the sum of between-clusters dispersion with the within-cluster dispersion

Empower Results®

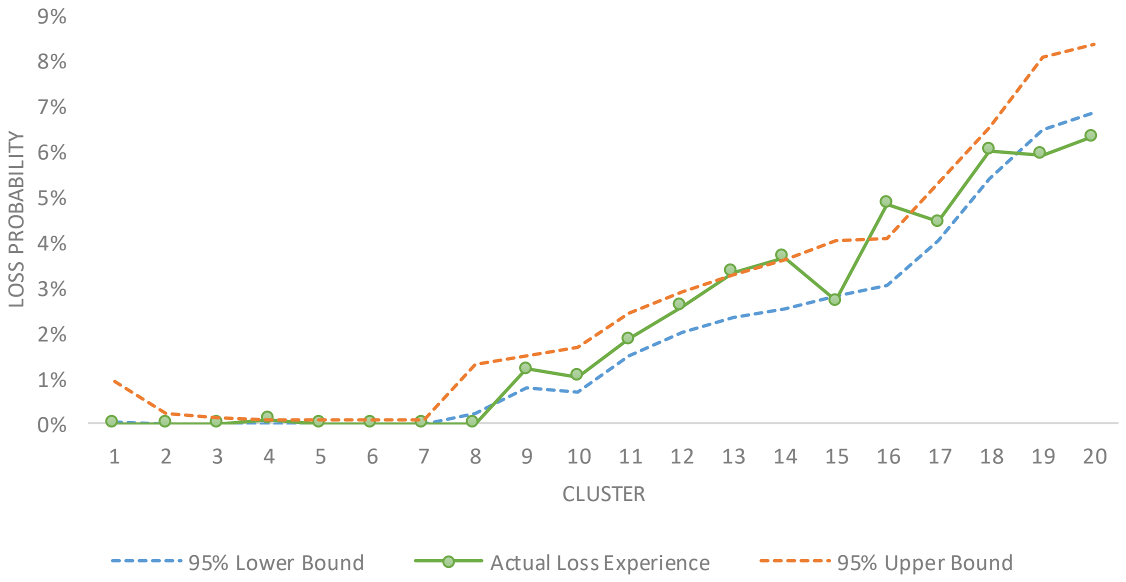
Unsupervised Clustering – Evaluations

	Clustering Stability Internal Measures on X's			St.d. (Estimated Loss Probability) $\sigma(\hat{p})$	Out-of-sample Forecasting Validation on Y		
	Davies–Bouldin Index	Silhouette Coefficient	Calinski Harabasz Score		Target Homogeneity σ(Actual Loss Count)	ρ(Estimated Loss Probability and the Actual Experience)	Avg Pricing Error / Over Price / Under Price
K-Means	2.30	0.09	2401	**0.1826%**	**22.13**	**97.25%**	0.029% / 0.198% / 0.169%
HAC*	**2.24**	**0.10**	**2459**	0.2056%	22.21	96.39%	**0.017% / 0.173% / 0.156%**

* Hierarchical Agglomerative Clustering

- All clustering algorithms above divide the total 100,000 risks into 20 clusters / risk cohorts to forecast risks at a macro level.
- After clustering, apply stratified random sampling to the clustered data points. 80% of the sample is treated as the training set for estimating the loss probability of each risk cohort and the remaining 20% is for validation purposes.

Unsupervised Clustering – Risk Prediction with K-Means

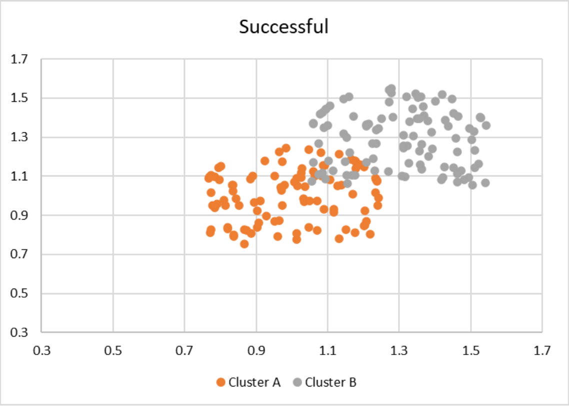


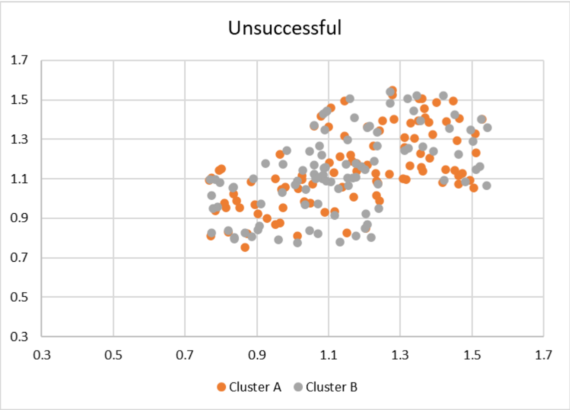
Cluster #	Estimated Loss Probability	Actual Loss Experience
#1	0.00%	0.00%
#2	0.00%	0.00%
#3	0.00%	0.00%
#4	0.00%	0.09%
#5	0.00%	0.00%
#6	0.00%	0.00%
#7	0.00%	0.00%
#8	0.60%	0.00%
#9	1.08%	1.21%
#10	1.09%	1.03%
#11	1.90%	1.85%
#12	2.43%	2.59%
#13	2.77%	3.34%
#14	3.02%	3.66%
#15	3.37%	2.70%
#16	3.54%	4.85%
#17	4.61%	4.44%
#18	5.94%	6.01%
#19	7.24%	5.93%
#20	7.57%	6.32%

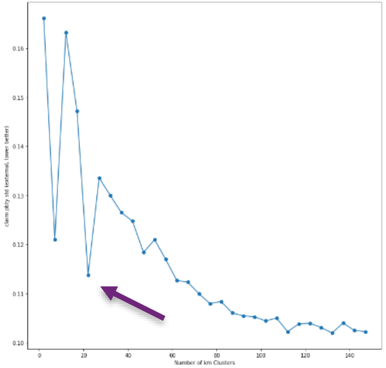
- The left compares the actual experience against the Wilson score confidence interval for true loss probability p (based on the assumption of the binomial distribution with continuity correction).
- On the right, the correlation ρ(estimated loss probability and the actual experience) = 0.9725.

Empower Results®

Unsupervised Clustering – Cluster Validation

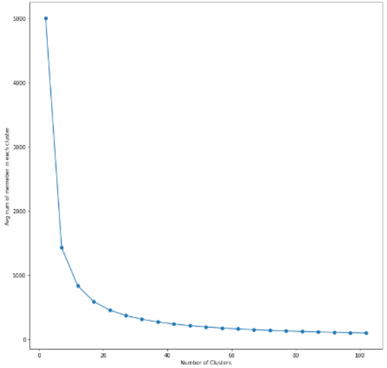






Standard Deviation of the Target Value

evaluates the average stability of the target variable (e.g. loss probability) within each cluster



Creditability Measure

counts the average number of sample points in each cluster

Empower Results®

Agenda

Profit Maximization

$$\max_{r_i} \sum_i prob(r_i, c_i)(r_i - c_i)$$

Where

i: auto insurance policy

r_i: auto insurance premium

c_i: auto insurance cost

$prob(r_i, c_i)$: probability that given r_i, the chances that policy i will be acquired or retained.

Possible constraints:

$c_i \leq C_{max}$	*underwriting rule*
$r_i - c_i \geq 0.02c_i$	*minimum profit requirement*
$\sum_i c_i \leq 50 \times Available\ Capital$	*capital sufficiency*

Improved Accuracy by Predictive Modeling

c_i

- More accurate estimation of auto insurance cost
- Fairer price

$prob(r_i, c_i)$

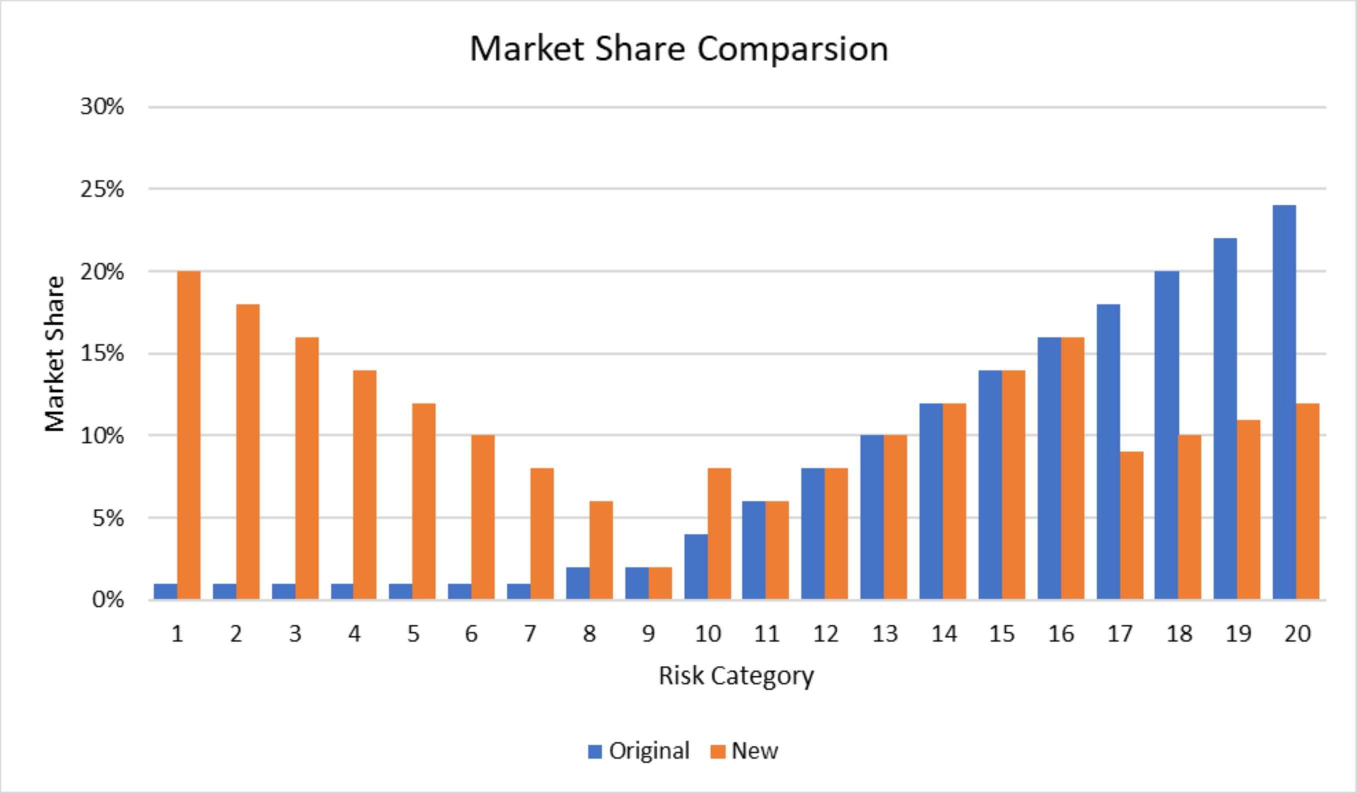
- Chances of retaining a policy
- Chances of wining a new policy

Impact of Improved Accuracy on Profit Maximization

			Original Strategy			New Strategy with Predictive Modeling		
Risk Category #	Loss Probability	# of Drivers (Entire Market)	Unit Profit	# of Policies	Total Profit	Unit Profit	# of Policies	Total Profit
#1	0.00%	4,400	20	44	880	3	880	2,640
#2	0.00%	12,150	19	122	2,309	3.5	2,187	7,655
#3	0.00%	24,450	18	245	4,401	4	3,912	15,648
#4	0.00%	40,550	17	406	6,894	4.5	5,677	25,547
#5	0.00%	62,800	16	628	10,048	5	7,536	37,680
#6	0.00%	73,800	15	738	11,070	5.5	7,380	40,590
#7	0.00%	84,600	14	846	11,844	6	6,768	40,608
#8	0.60%	57,550	13	1,151	14,963	6.5	3,453	22,445
#9	1.08%	24,250	12	485	5,820	7	728	5,093
#10	1.09%	49,750	11	1,990	21,890	7.5	995	7,463
#11	1.90%	43,200	10	2,592	25,920	8	3,024	24,192
#12	2.43%	54,100	9	4,328	38,952	8.5	4,599	39,087
#13	2.77%	46,250	8	4,625	37,000	8	4,625	37,000
#14	3.02%	53,900	7	6,468	45,276	7	6,468	45,276
#15	3.37%	45,100	6	6,314	37,884	6	6,314	37,884
#16	3.54%	58,550	5	9,368	46,840	5	9,368	46,840
#17	4.61%	63,950	4	11,511	46,044	5	5,756	28,778
#18	5.94%	54,800	3	10,960	32,880	5	5,480	27,400
#19	7.24%	86,500	2	19,030	38,060	5	9,515	47,575
#20	7.57%	59,350	1	14,244	14,244	5	7,122	35,610
Average/**Total**	2.67%	**1,000,000**	4.7	**96,094**	**453,218**	5.6	**101,786**	**575,008**

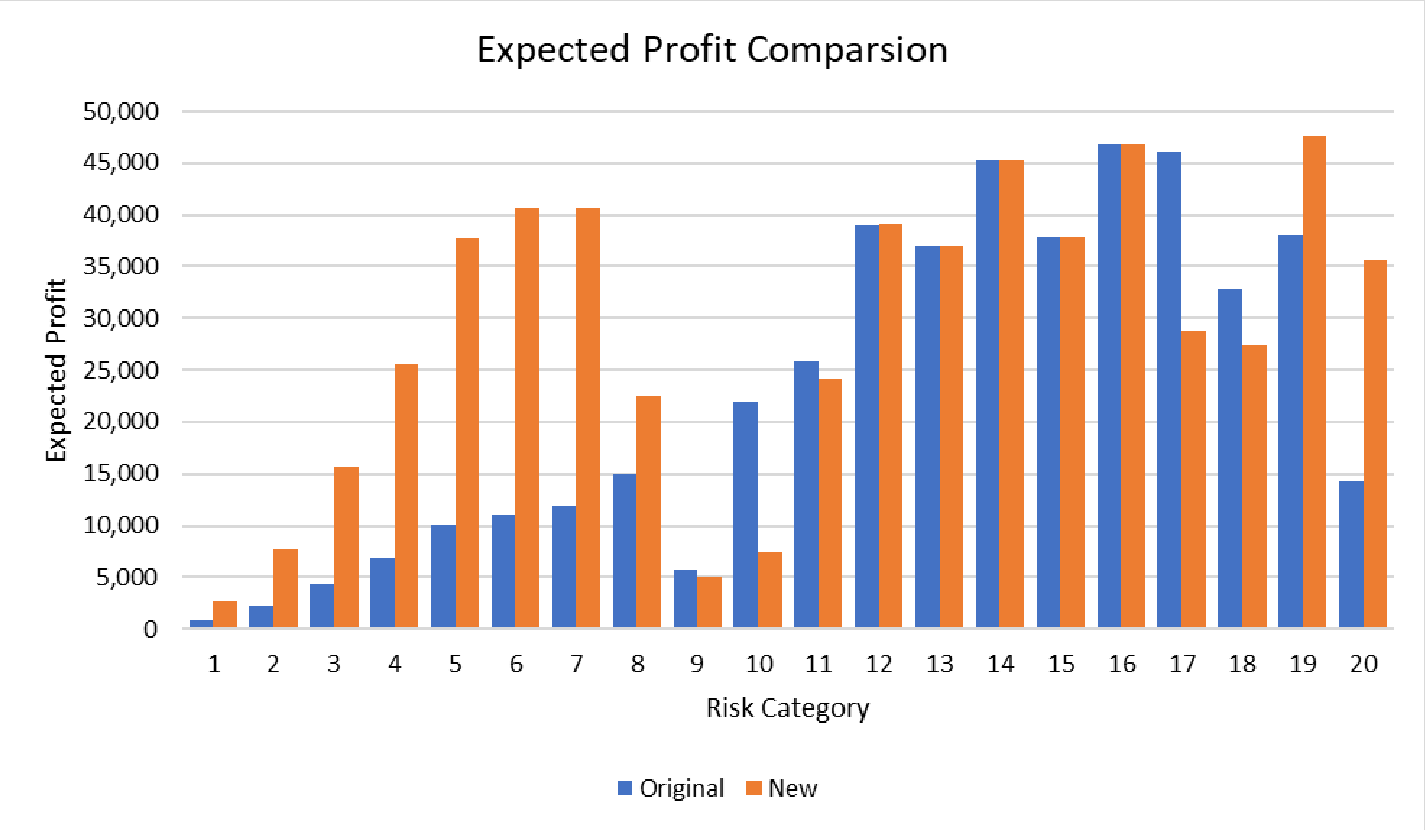
Illustration only

Impact of Predictive Modeling – Market Share



Shifting from high risk profile to low risk profile

Impact of Predictive Modeling – Expected Profit



Higher total profit due to low risks

Empower Results®

Impact of Predictive Modeling – Summary

	Original	New
Market Share	9.6%	10.2%
Average Profit	$4.7	$5.6
Total Profit	$0.45 Mil	$0.58 Mil
Return on Risk Adjusted Capital	6.8%	11.0%

Considerations

- Predictive modeling enables more accurate profit maximization through not only the cost estimation but also the improved demand function.
- Market dynamics requires model recalibration from time to time. Automated process allows timely decision-making.
- To ensure a fair process, predictive models should use the same set of predictors for all policies to meet regulatory requirements.

Agenda

Section 1	Introduction to Predictive Modeling
Section 2	Applying Predictive Modeling to Auto Insurance Pricing
Section 3	Pricing Optimization with Better Prediction
Section 4	Recap and Q&A

Recap

- Predictive modeling can improve prediction accuracy
- To achieve optimal performance, a robust process is needed to fine-tune and validate models
- Model training can be parallelized to ensure timely delivery
- Need an integrated platform to perform advanced predictive analytics
- Need expertise for data preparation, model validation, model selection and result interpretation
- Need to update models as business conditions and circumstance require
- Predictive modeling can be applied to risk management and improving business efficiency

Thank you!

Q&A

Thank You

Contact the speakers :

Peter Phillips
416.598.7133
peter.phillips@aon.com

Kevin Shang
416.263.7908
kevin.shang@aon.com

Aon
Empower Results®