SEAC Fall 2022

Leading Your Organization in Machine Learning

Aaron Schaffer, ASA, MAAA

Leading Your Organization in Machine Learning

01

Overview of Machine Learning

- What is ML?
- Big Picture
- Examples

02

Why You Should Embrace ML as a Leader

- ASA Requirements
- Team Efficiencies
- Speed to Insight
- Competitive Advantages

03

Common Solutions for Actuarial Work

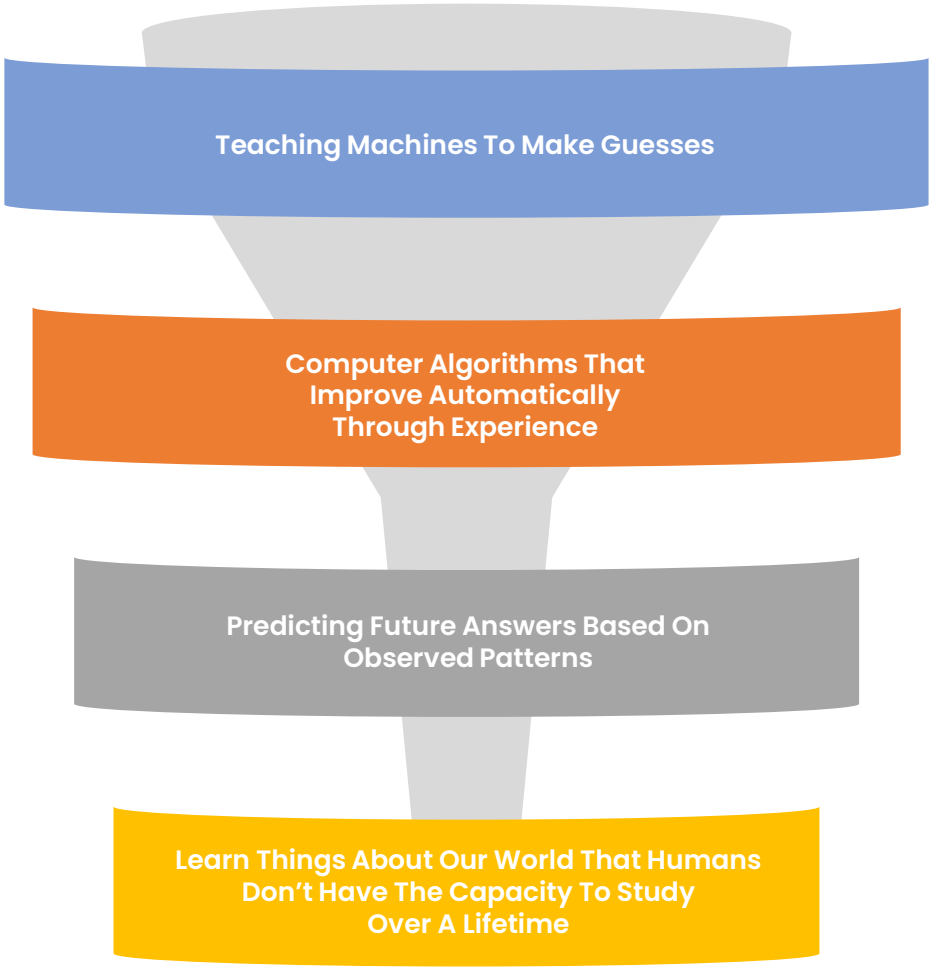
- Forecasts & Predictions
- Segmentation
- Data Imputation
- Outlier Analysis
- KPI Drivers

04

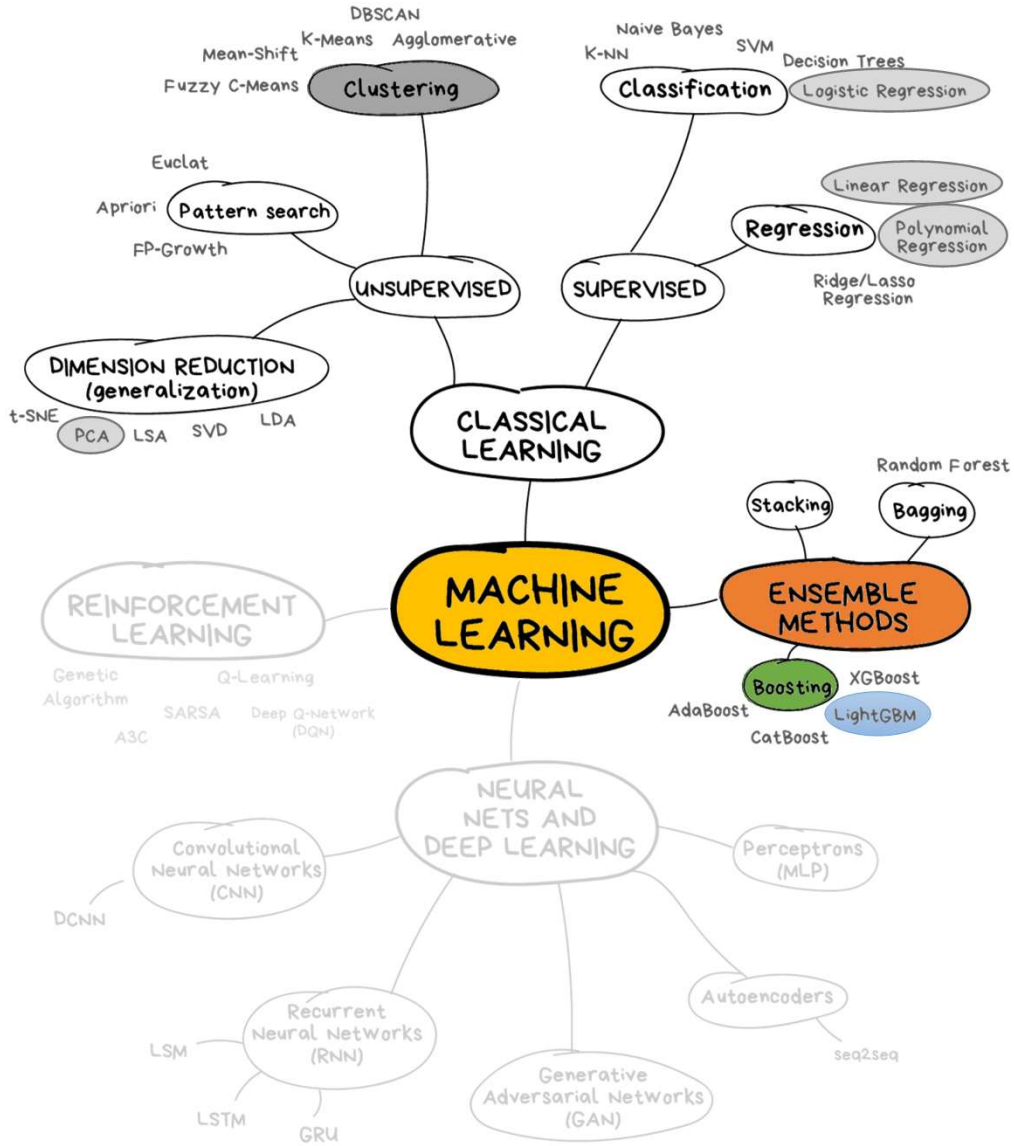
How to Get Your Team Started

- Tools and Applications
- Skills and Competencies
- Online Resources
- Continuing Ed

What is Machine Learning?

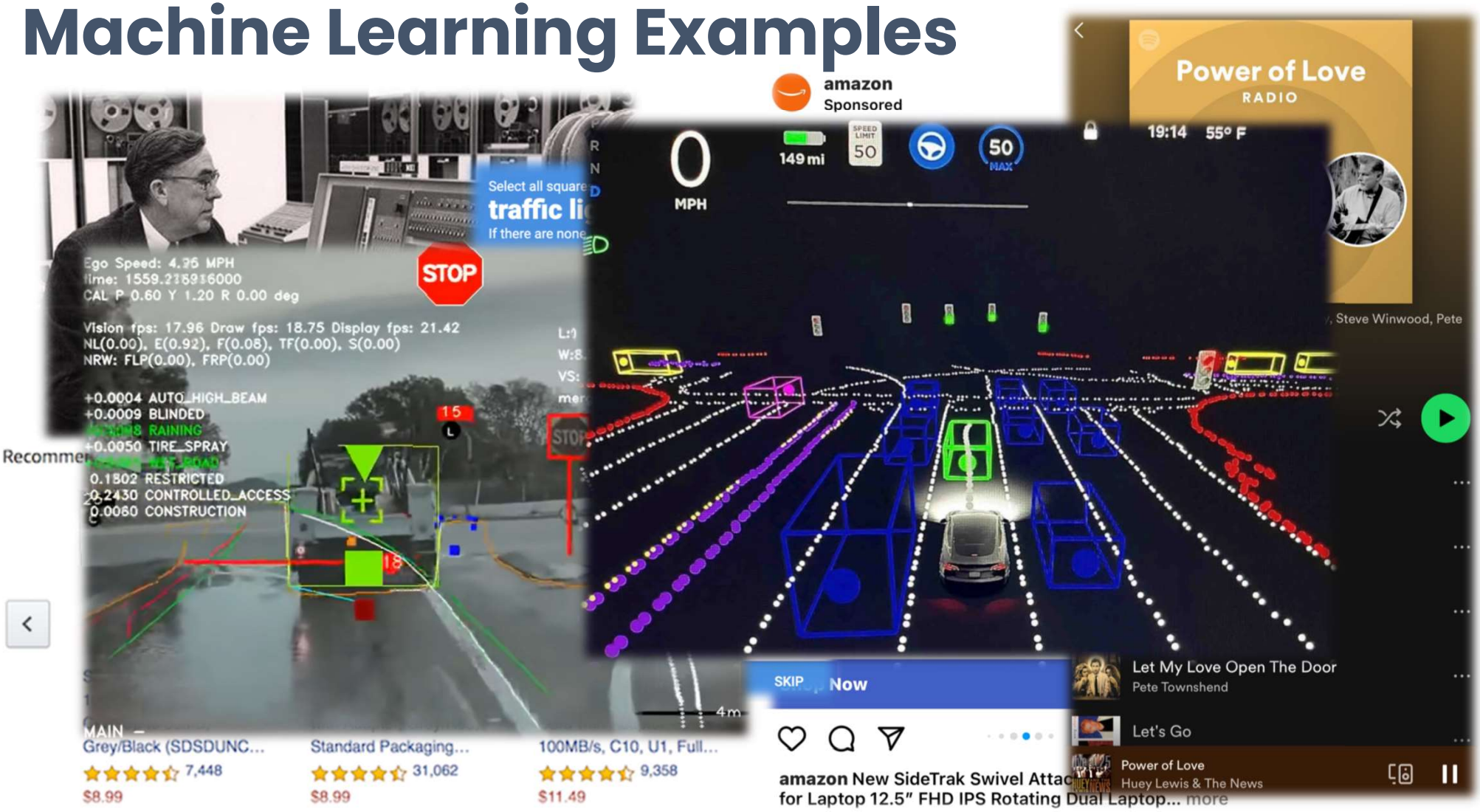


The. Big. Picture.



https://vas3k.com/blog/machine_learning/

Machine Learning Examples



Leading Your Organization in Machine Learning

01

Overview of Machine Learning

- What is ML?
- Big Picture
- Examples

02

Why You Should Embrace ML as a Leader

- ASA Requirements
- Team Efficiencies
- Speed to Insight
- Easy to Implement

03

Common Solutions for Actuarial Work

- Forecasts & Predictions
- Segmentation
- Data Imputation
- Outlier Analysis
- KPI Drivers

04

How to Get Your Team Started

- Tools and Applications
- Skills and Competencies
- Online Resources
- Continuing Ed

Why You Should Embrace ML as a Leader

Starting with the class of 2022…

New ASAs will have passed 2 exams and 1 module where the syllabus focuses primarily on data science and machine learning.

Exams:
Statistics for Risk Modeling (30%)
Predictive Analytics (75%)

Module:
Advanced Topics in Predictive Analytics (100%)

https://www.soa.org/49926f/globalassets/assets/files/edu/2022/2022-09-exam-srm-syllabus.pdf
https://www.soa.org/4a9f06/globalassets/assets/files/edu/2023/2022-04-exam-pa-syllabus.pdf
https://www.soa.org/48d54a/globalassets/assets/files/edu/2022/2022-01-atpa-learning-objectives.pdf

Why You Should Embrace ML as a Leader

Save Time

Build a framework.
Reduce time needed to analyze data compared to traditional methods.
High quality results.

Speed

Outperform

Get the value.
ML model results from a small team can rival the output and efforts of a large analytical unit.

Efficiency

Test & Learn

Grow new areas.
Research drivers of KPIs with a lower hurdle rate to stand up a new team or project.

Invest

Find an Edge

Stay relevant.
Companies are looking for advantages using ML while learning into best practices and techniques

Competition

Leading Your Organization in Machine Learning

01

Overview of Machine Learning

- What is ML?
- Big Picture
- Examples

02

Why You Should Embrace ML as a Leader

- ASA Requirements
- Team Efficiencies
- Speed to Insight
- Easy to Implement

03

Common Solutions for Actuarial Work

- Forecasts & Predictions
- Segmentation
- Data Imputation
- Outlier Analysis
- KPI Drivers

04

How to Get Your Team Started

- Tools and Applications
- Skills and Competencies
- Online Resources
- Continuing Ed

Forecasts & Predictions

Common Solutions for Actuarial Work

Difficulty: Advanced

A

Predict Y using $A_1 \ldots Z_n$

Claims, Utilization, Visits, Trips, Scripts, Sales, Conditions, ...

B

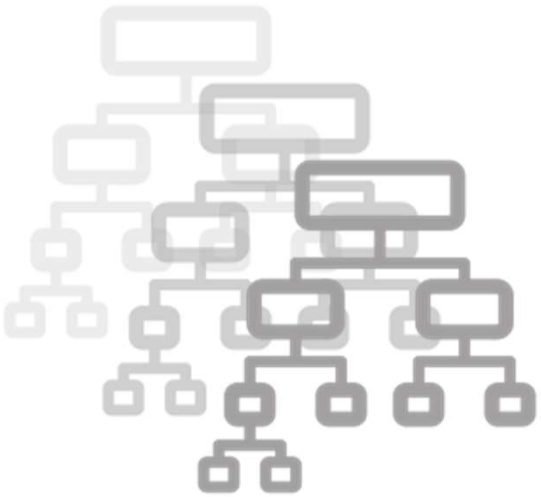
Train & Validate

Feed in large training dataset, refine model parameters, control output for overfitting

C

Solution: LightGBM

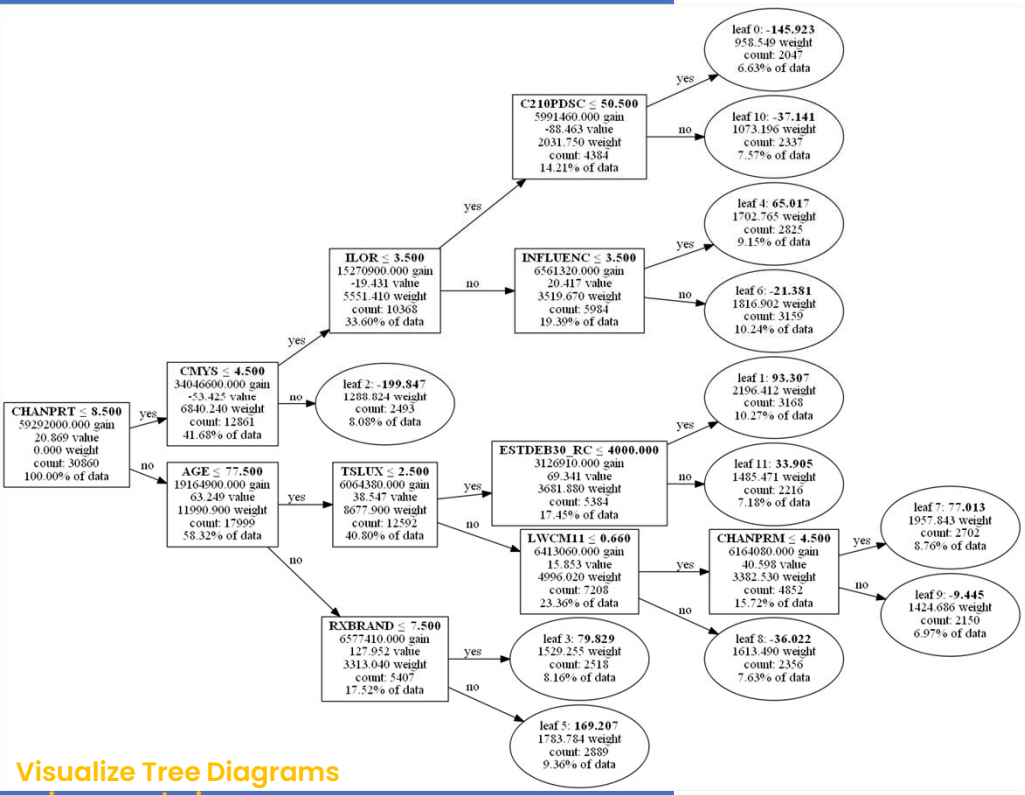
Very accurate, runs quickly, supports continuous and categorical features, parameters for multi-use optimization



Gradient Boosted Decision Trees

https://lightgbm.readthedocs.io/en/v3.3.3/

What Is A Gradient Boosted Decision Tree?



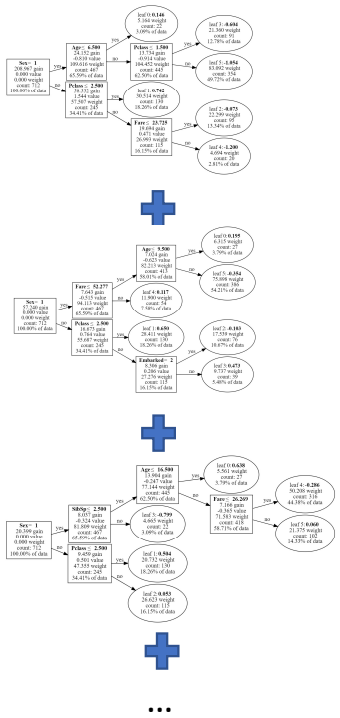
Visualize Tree Diagrams using graphviz

- A model that creates an algorithm consisting of ensembled decision trees to fit your data
- Multiple decision trees (usually 100s) are computed and then ensembled together using a learning rate factor
- 1st tree is fit to your data based on the mean of your prediction target

What Is A Gradient Boosted Decision Tree?

```
Starting training...
[1]     valid_0's rmse: 0.408328
Training until validation scores don't improve for 5 rounds
[2]     valid_0's rmse: 0.38368
[3]     valid_0's rmse: 0.373163
[4]     valid_0's rmse: 0.371203
[5]     valid_0's rmse: 0.367234
[6]     valid_0's rmse: 0.368405
[7]     valid_0's rmse: 0.367496
[8]     valid_0's rmse: 0.363573
[9]     valid_0's rmse: 0.363874
[10]    valid_0's rmse: 0.364973
[11]    valid_0's rmse: 0.365873
[12]    valid_0's rmse: 0.364953
[13]    valid_0's rmse: 0.362096
[14]    valid_0's rmse: 0.362784
[15]    valid_0's rmse: 0.359469
[16]    valid_0's rmse: 0.36013
[17]    valid_0's rmse: 0.361306
[18]    valid_0's rmse: 0.362816
[19]    valid_0's rmse: 0.362301
[20]    valid_0's rmse: 0.360433
Early stopping, best iteration is:
[15]    valid_0's rmse: 0.359469
```

Training Round | Error measure | Error

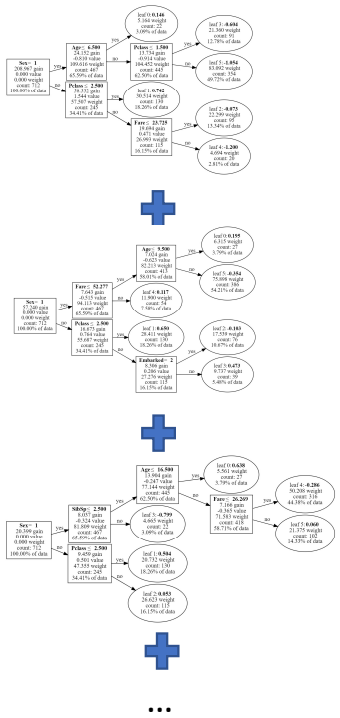


- 2nd tree fits to the residuals produced by 1st tree
  - All subsequent trees are fit to the residuals from the previous tree
- Parent nodes, child nodes, and split points are determined by the model using measures like entropy, information gain, and squared errors
- Typically use a train-test dataset split of 75% / 25%
  - Optional cross-validation to partition dataset and create separate models from each partition

What Is A Gradient Boosted Decision Tree?

```
Starting training...
[1]     valid_0's rmse: 0.408328
Training until validation scores don't improve for 5 rounds
[2]     valid_0's rmse: 0.38368
[3]     valid_0's rmse: 0.373163
[4]     valid_0's rmse: 0.371203
[5]     valid_0's rmse: 0.367234
[6]     valid_0's rmse: 0.368405
[7]     valid_0's rmse: 0.367496
[8]     valid_0's rmse: 0.363573
[9]     valid_0's rmse: 0.363874
[10]    valid_0's rmse: 0.364973
[11]    valid_0's rmse: 0.365873
[12]    valid_0's rmse: 0.364953
[13]    valid_0's rmse: 0.362096
[14]    valid_0's rmse: 0.362784
[15]    valid_0's rmse: 0.359469
[16]    valid_0's rmse: 0.36013
[17]    valid_0's rmse: 0.361306
[18]    valid_0's rmse: 0.362816
[19]    valid_0's rmse: 0.362301
[20]    valid_0's rmse: 0.360433
Early stopping, best iteration is:
[15]    valid_0's rmse: 0.359469
```

Training Round | Error measure | Error



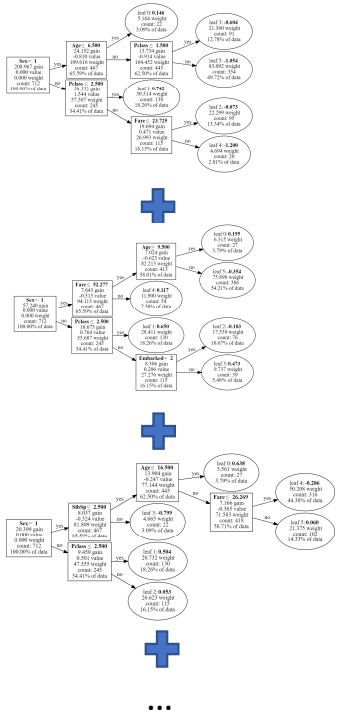
- Training rounds stop when the Test residuals begin to increase rather than decrease
- Hyperparameters are tuned to control for overfitting balanced with accuracy of the final model
- Compute R^2, decile chart, or AUC to evaluate model fit

Tune the parameters and re-train to improve accuracy

What Is A Gradient Boosted Decision Tree?

```
Starting training...
[1]     valid_0's rmse: 0.408328
Training until validation scores don't improve for 5 rounds
[2]     valid_0's rmse: 0.38368
[3]     valid_0's rmse: 0.373163
[4]     valid_0's rmse: 0.371203
[5]     valid_0's rmse: 0.367234
[6]     valid_0's rmse: 0.368405
[7]     valid_0's rmse: 0.367496
[8]     valid_0's rmse: 0.363573
[9]     valid_0's rmse: 0.363874
[10]    valid_0's rmse: 0.364973
[11]    valid_0's rmse: 0.365873
[12]    valid_0's rmse: 0.364953
[13]    valid_0's rmse: 0.362096
[14]    valid_0's rmse: 0.362784
[15]    valid_0's rmse: 0.359469
[16]    valid_0's rmse: 0.36013
[17]    valid_0's rmse: 0.361306
[18]    valid_0's rmse: 0.362816
[19]    valid_0's rmse: 0.362301
[20]    valid_0's rmse: 0.360433
Early stopping, best iteration is:
[15]    valid_0's rmse: 0.359469
```

Training Round | Error measure | Error



The concept is:

A single tree is a “weak learner”

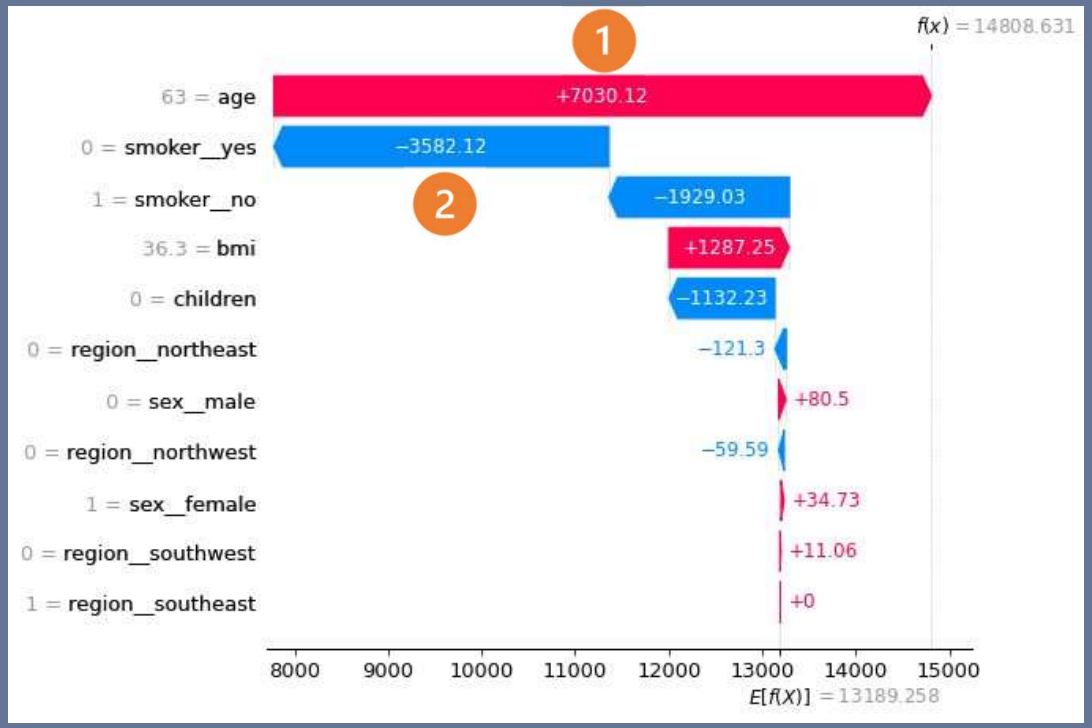
But an ensemble of numerous trees together produces a “strong learner”

Questions?

Feature Importance: Shapley Values

Shapley Waterfall

Prediction for Individual



Average of all Claims

Shapley Values quantify marginal contribution of a feature on the prediction…

…and are relative to the average of the predicted target

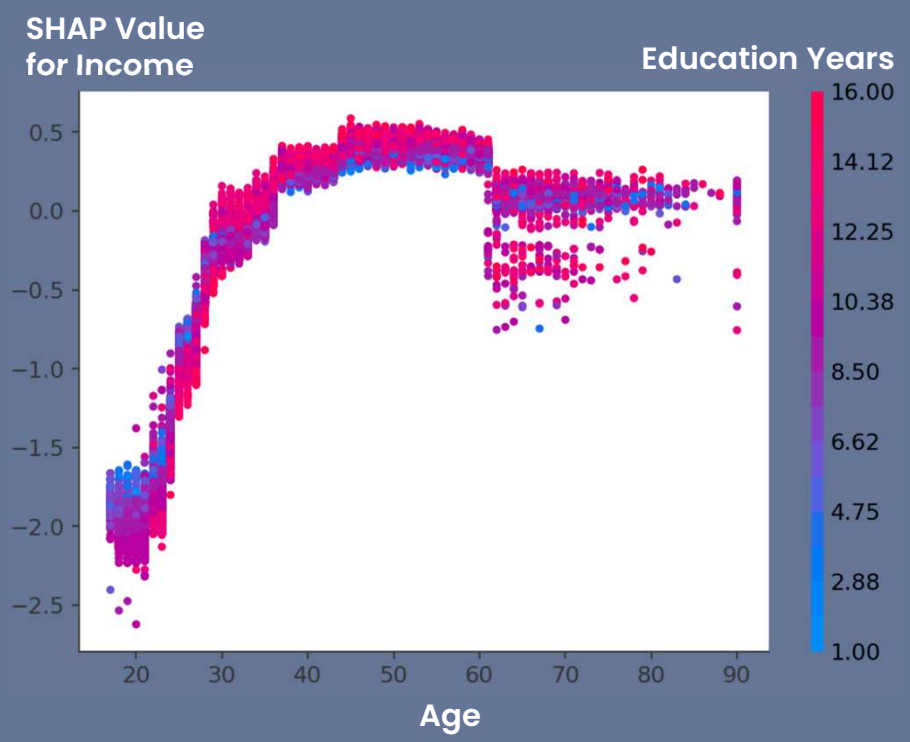
1. Age = 63 increases the claim prediction by $7k
2. Smoker = No decreases the claim prediction by $3.6k

https://www.actuaries.digital/2021/02/05/explainable-ml-a-peek-into-the-black-box-through-shap/

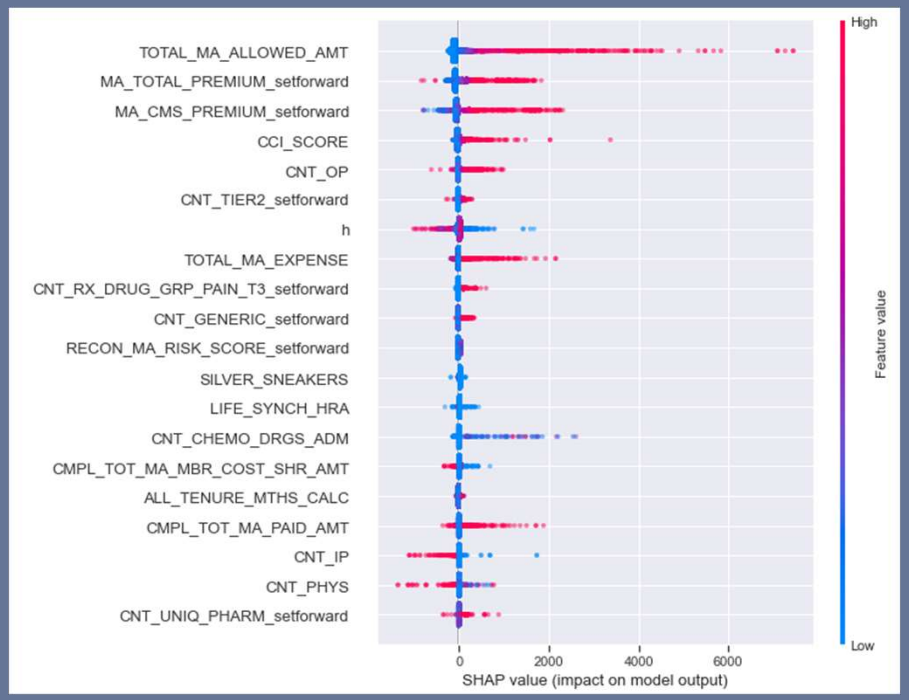
Feature Importance: Shapley Values

Run the SHAP package on your model to understand impact of features on the predictions

Dependence Plot



Beeswarm Plot



https://shap.readthedocs.io/en/latest/index.html

Segmentation

Common Solutions for Actuarial Work

Difficulty: Intermediate

A

Cohort Analysis

Find the most impactful characteristics that maximize classification value among groups

B

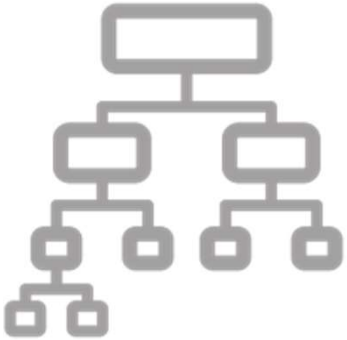
Train & Review

Feed in large training dataset, refine model parameters, focus on leaf credibility

C

Solution: LightGBM

Natively supports continuous and categorical features, runs quickly, parameters for multi-use optimization



Decision Tree

https://lightgbm.readthedocs.io/en/v3.3.3/

Outlier Analysis

Common Solutions for Actuarial Work

Difficulty: Intermediate

A

Anomaly Detection

Find outliers with multiple dimensions of features

B

Unsupervised

Not necessary to label or structure data, recommend high performance PC for large datasets

C

Solution: Extended Isolation Forest

Finds local outliers—not just corner cases, create anomaly scores, refine contamination rate, visualize results

Extended Isolation Forest

https://github.com/sahandha/eif

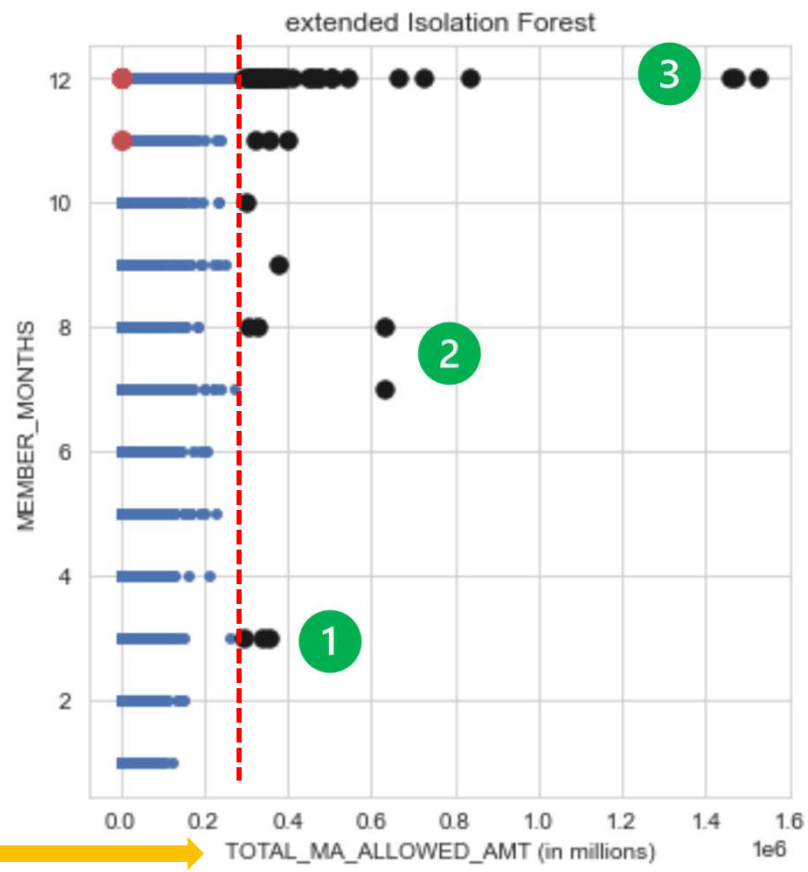
Outlier Analysis

Common Solutions for Actuarial Work

Difficulty: Intermediate

Dimension #1
Member Months

Dimension #2
Allowed Claims Cost



Conducting a Power Analysis but the resulting sample size was too large...

...over 150k members required and business would not fund that approach...

...recommended removing outliers to reduce population variance and therefore reduce required sample size significantly

Outlier Analysis

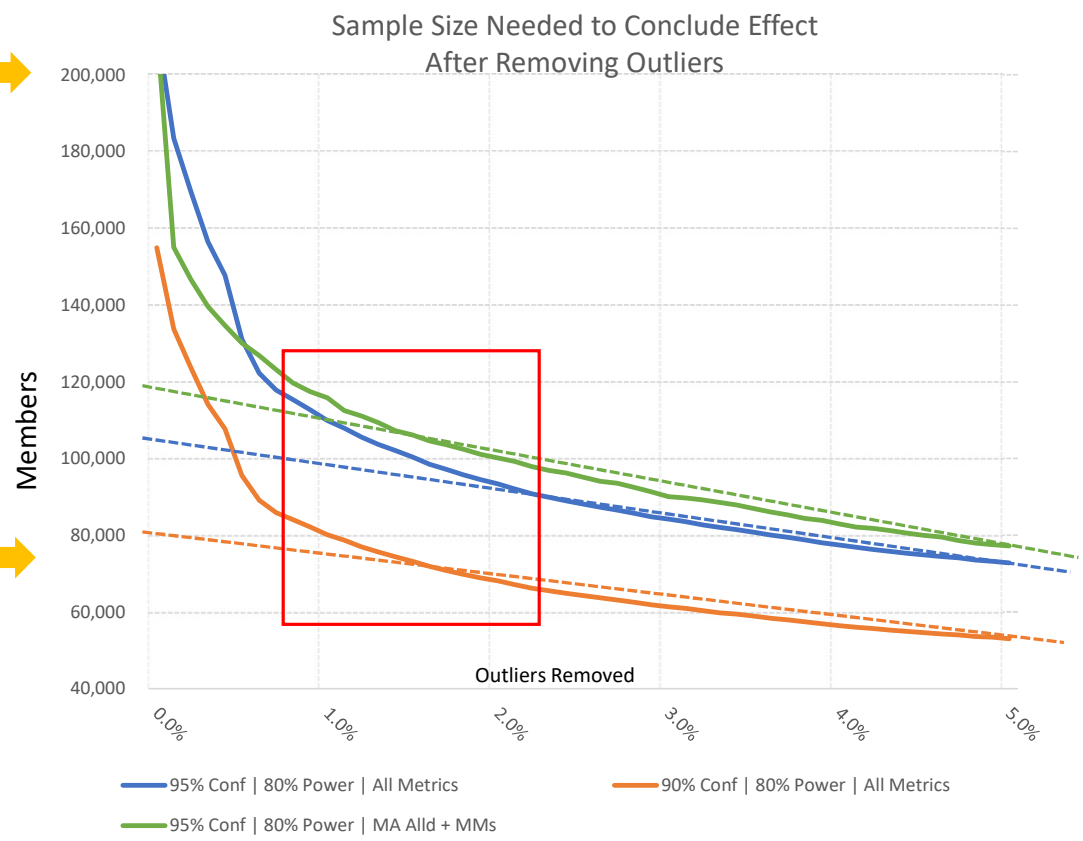
Common Solutions for Actuarial Work

Difficulty: Intermediate

+150k Sample Size Unacceptable ...Too Large

Remove 1.5% as Outliers

Business Acceptable Sample Size



Key Influencers & Top Segments

Common Solutions for Actuarial Work

Difficulty: Simple

A

Linear & Logistic Regressions

Analyze factors that impact your dataset, clearly understand drivers in simple terms

B

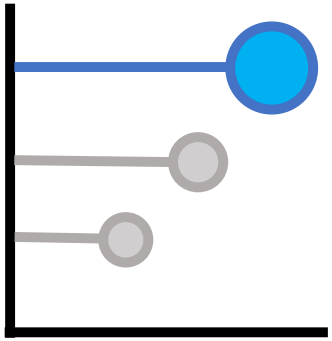
KPI Drivers

Import dataset to Power BI, choose Analyze and Explain By variables, explore the results

C

Solution: Power BI Key Influencers

Fast to implement, easy to interpret, use categories as high-level drivers, clean your data for outliers and multicollinearity first



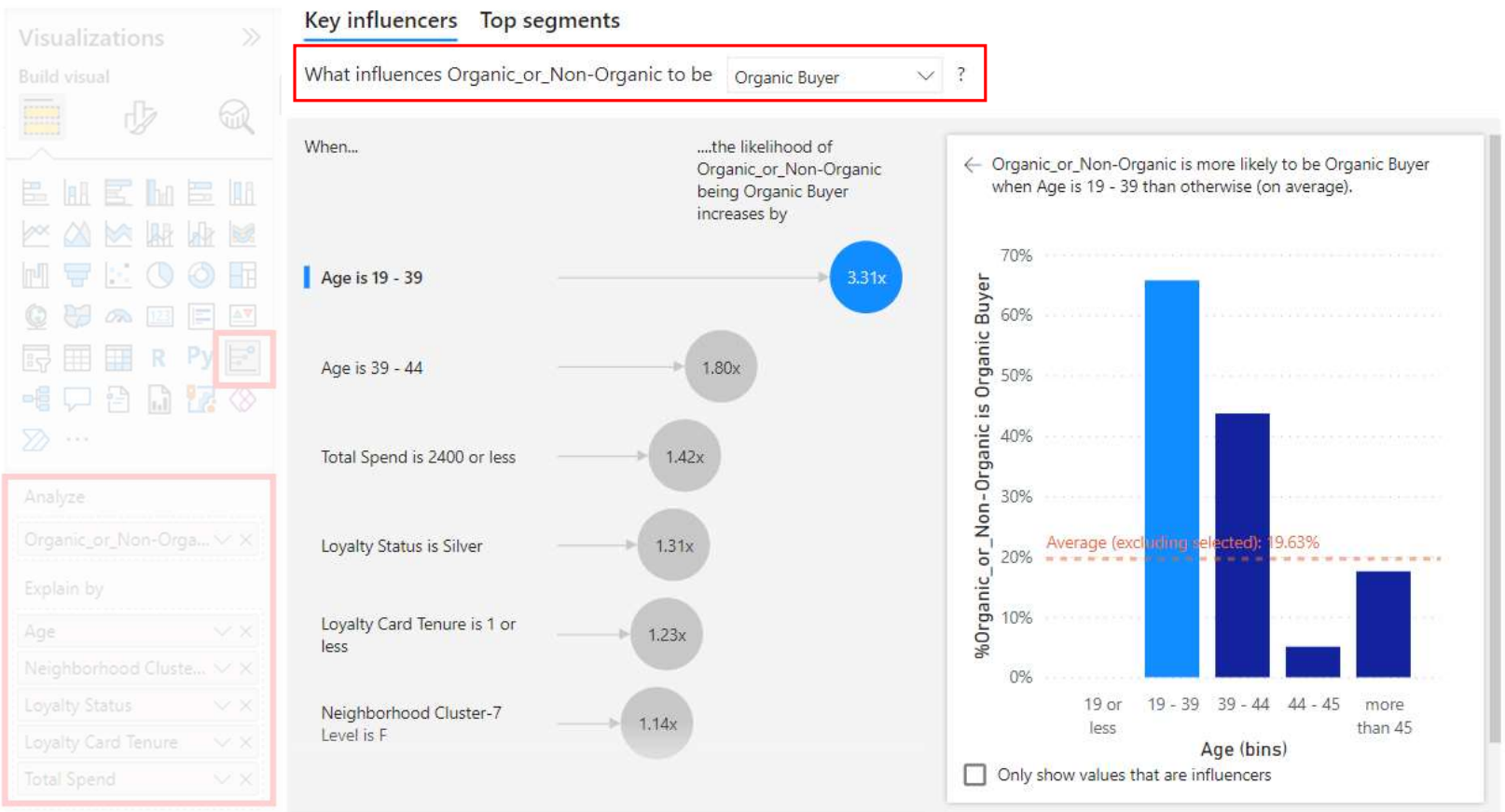
Power BI Key Influencers

https://www.bluegranite.com/blog/key-influencers-visualization-regression-models-in-power-bi

Key Influencers & Top Segments

Difficulty: Simple

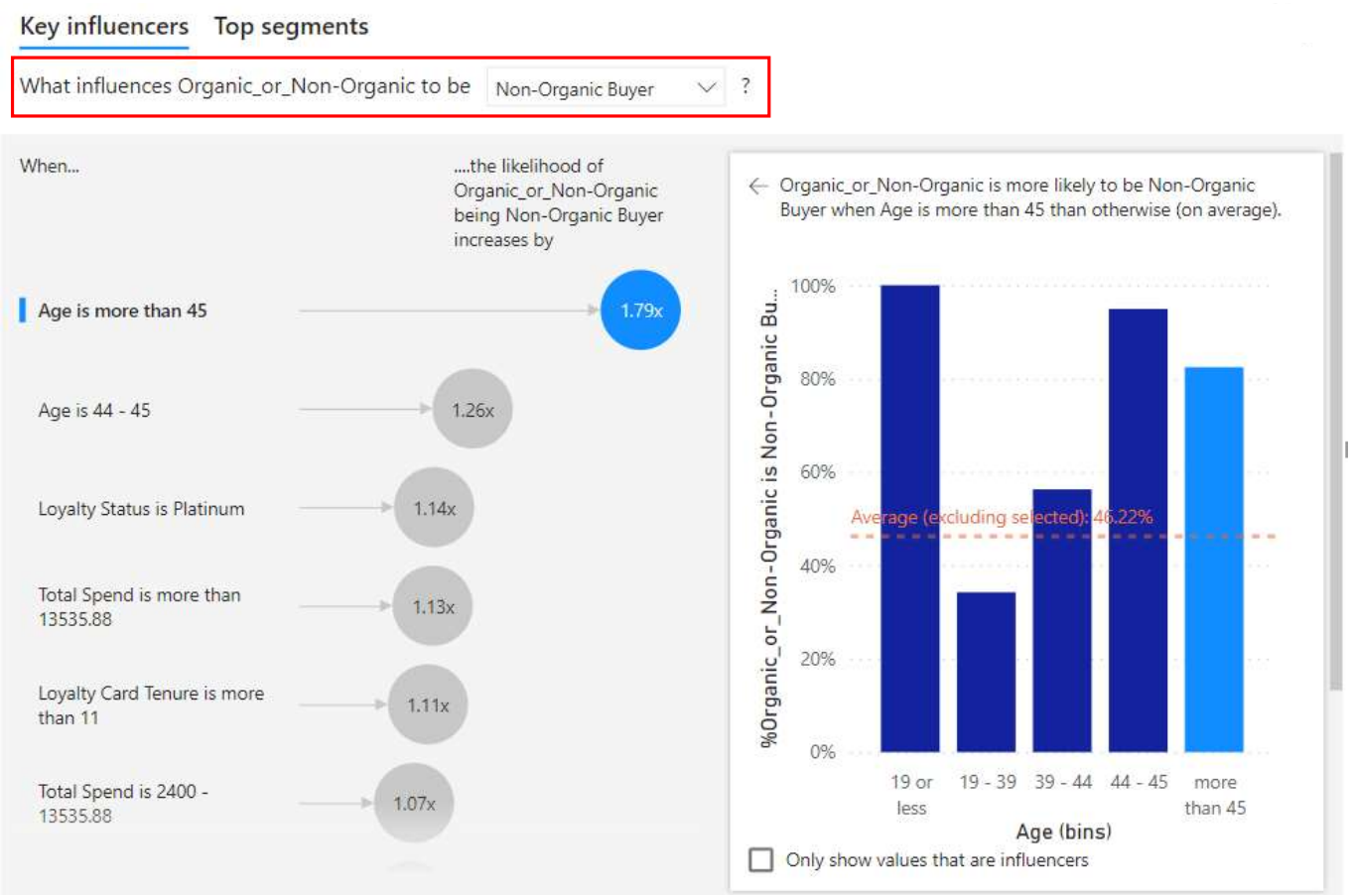
Common Solutions for Actuarial Work



Key Influencers & Top Segments

Difficulty: Simple

Common Solutions for Actuarial Work



Key Influencers & Top Segments

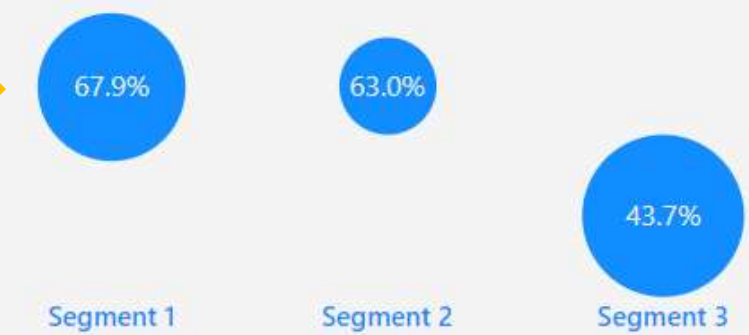
Common Solutions for Actuarial Work

Difficulty: Simple

Key influencers | Top segments

When is Organic_or_Non-Organic more likely to be Organic Buyer ?

We found 3 segments and ranked them by % Organic_or_Non-Organic is O...



	Segment 1	Segment 2	Segment 3
% Organic_or_Non-O...	67.9%	63.0%	43.7%
Population count	788	608	842

% of Segment
Where Condition = TRUE

Bubble Size
Relative to Total Population

Bonus

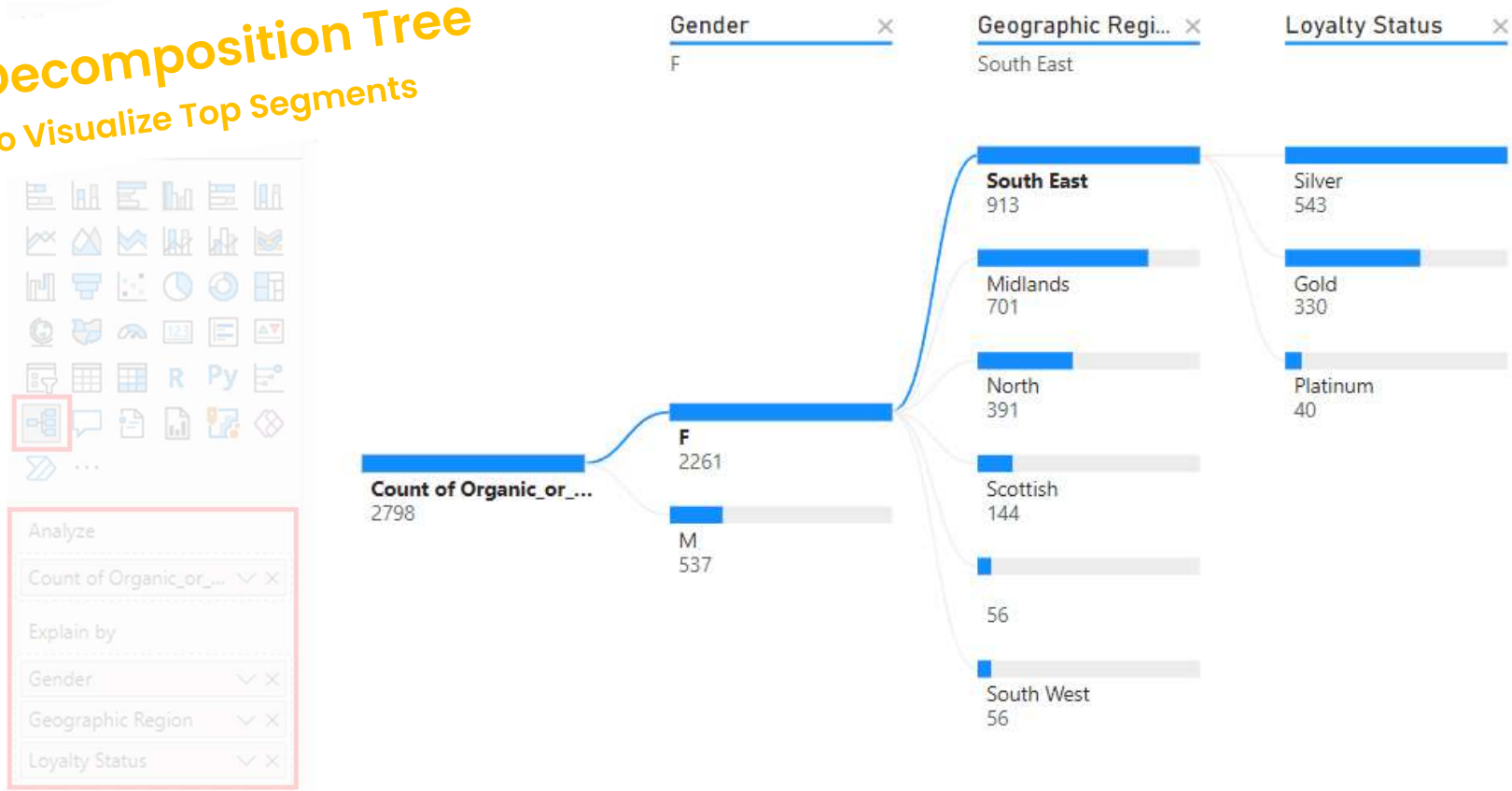
Key Influencers & Top Segments

Common Solutions for Actuarial Work

Difficulty: Simple

Decomposition Tree

to Visualize Top Segments



Leading Your Organization in Machine Learning

01
Overview of Machine Learning

- What is ML?
- Big Picture
- Examples

02
Why You Should Embrace ML as a Leader

- ASA Requirements
- Team Efficiencies
- Speed to Insight
- Easy to Implement

03
Common Solutions for Actuarial Work

- Forecasts & Predictions
- Segmentation
- Data Imputation
- Outlier Analysis
- KPI Drivers

04
How to Get Your Team Started

- Tools and Applications
- Skills and Competencies
- Online Resources
- Continuing Ed

How to Get Your Team Started

Discover and Decide Your Path

No Code

Power BI Key Influencers,
XLStat, XLMiner

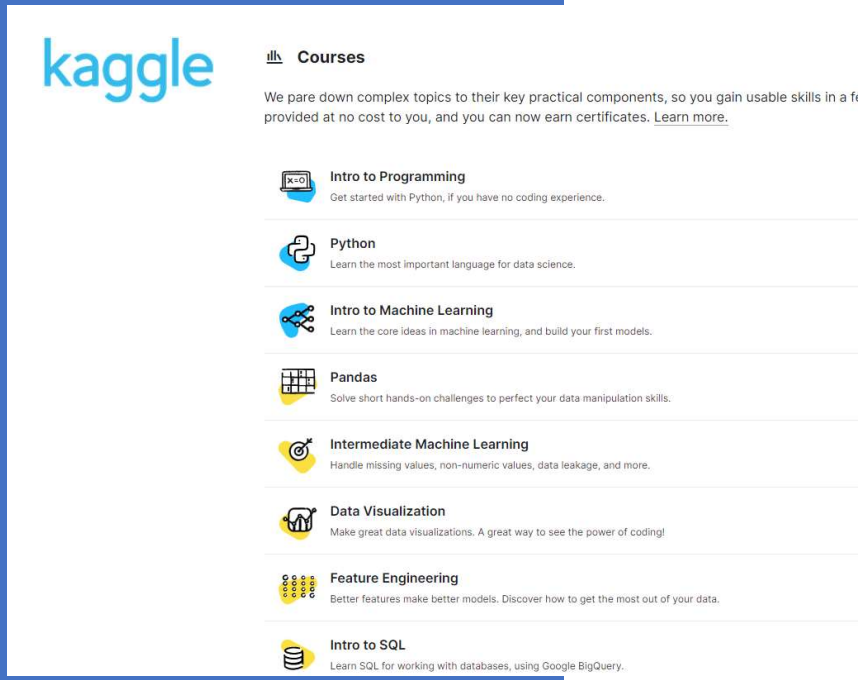
Code

Python or RStudio
PyCharm, JupyterLab
LightGBM, scikit-learn, ...

Low Code

Alteryx, DataRobot,
Python & PyCaret

How to Get Your Team Started



Be An Engaged Leader

- Sign up for every ML-oriented SOA and AAA webinar you see in your inbox
- Attend every ML session you can at SEAC and SOA meetings
- Watch YouTube videos on ML subjects

Check Out Kaggle's Learn

- Free courses that take hours, not months, to learn concepts. It's very visual and simple to follow along
- kaggle.com/learn

Adopt the Terminology

Remind yourself and your team this is a new frontier for actuarial work and learning to speak the language is important just like delivering the results

https://www.kaggle.com/learn