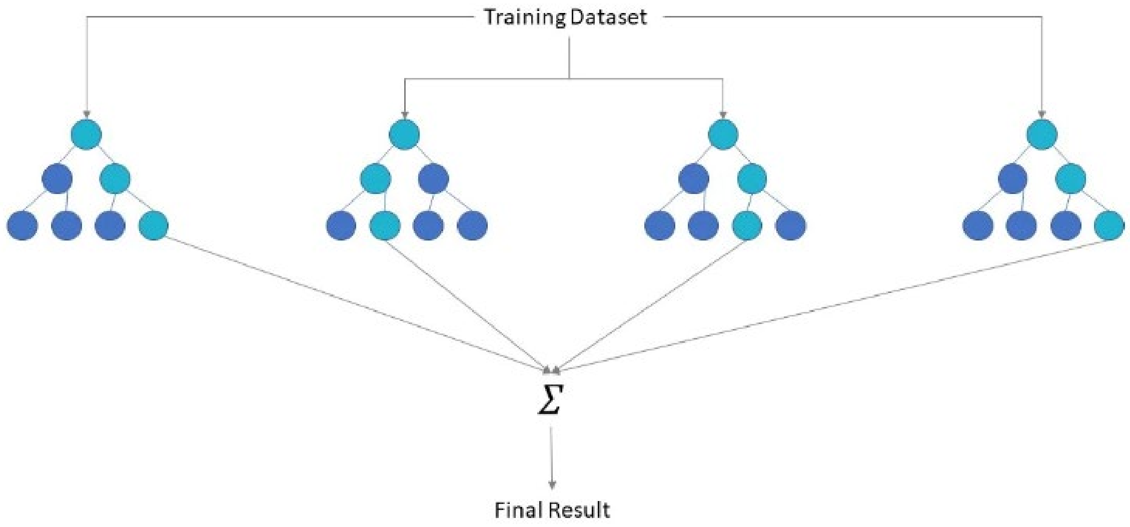


Random Forest Models

Sam Kloese, ACAS, CSPA
P/C Rate Modeling Actuary
December 21, 2021

NAIC NATIONAL ASSOCIATION OF INSURANCE COMMISSIONERS

Introduction

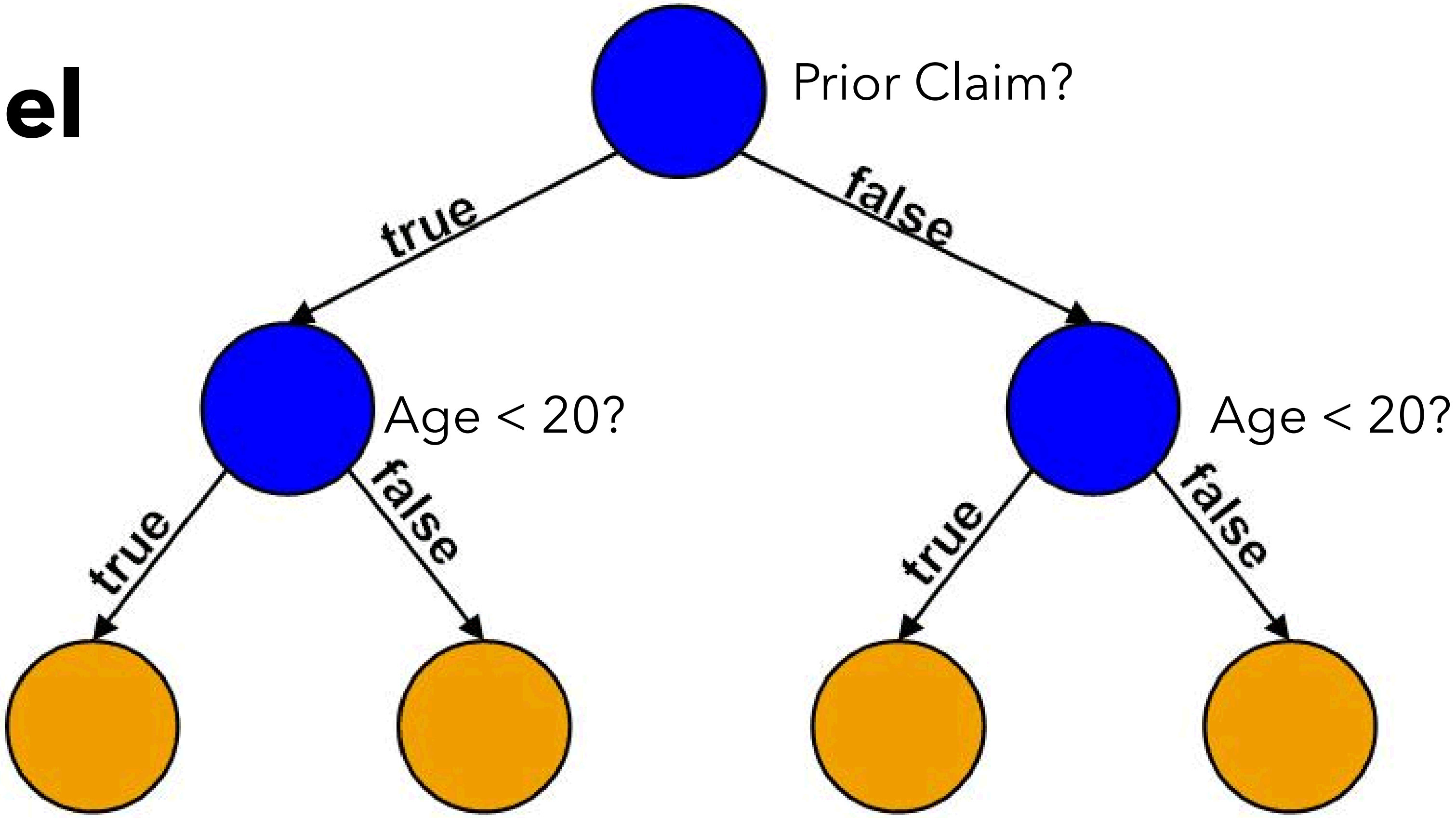
- GLM’s are industry standard
- The CASTF White Paper for Predictive Models is focused primarily on GLM’s
- Some companies are filing with more sophisticated models
  - GAM – Similar to GLM’s, but with non-parametric “smoothed” terms
  - Tree Based Models – Based on a collection of multiple decision trees
  - Neural Networks – Mostly for generating scores based on images
- The NAIC model review team has reviewed the above model types without CASTF guidance
- The NAIC model review team would like to discuss how reviews should vary for these differing model types
- Today’s focus is on Random Forests (a type of Tree Based Model)

Tree Based Models

- Models that can be represented as a decision tree or a collection of decision trees
- Types of Tree Based Models
  - Single decision Tree
  - “Bagged” Trees
  - Random Forest
  - Gradient Boosting Machine (XGBoost)
- Supervised Model
  - There is still a target variable
    - Classification: Renew/Non-renew, Claim/No Claim, Fraud/No Fraud
    - Regression: Frequency, Severity, Pure Premium
- Today’s focus will be on Random Forest Models

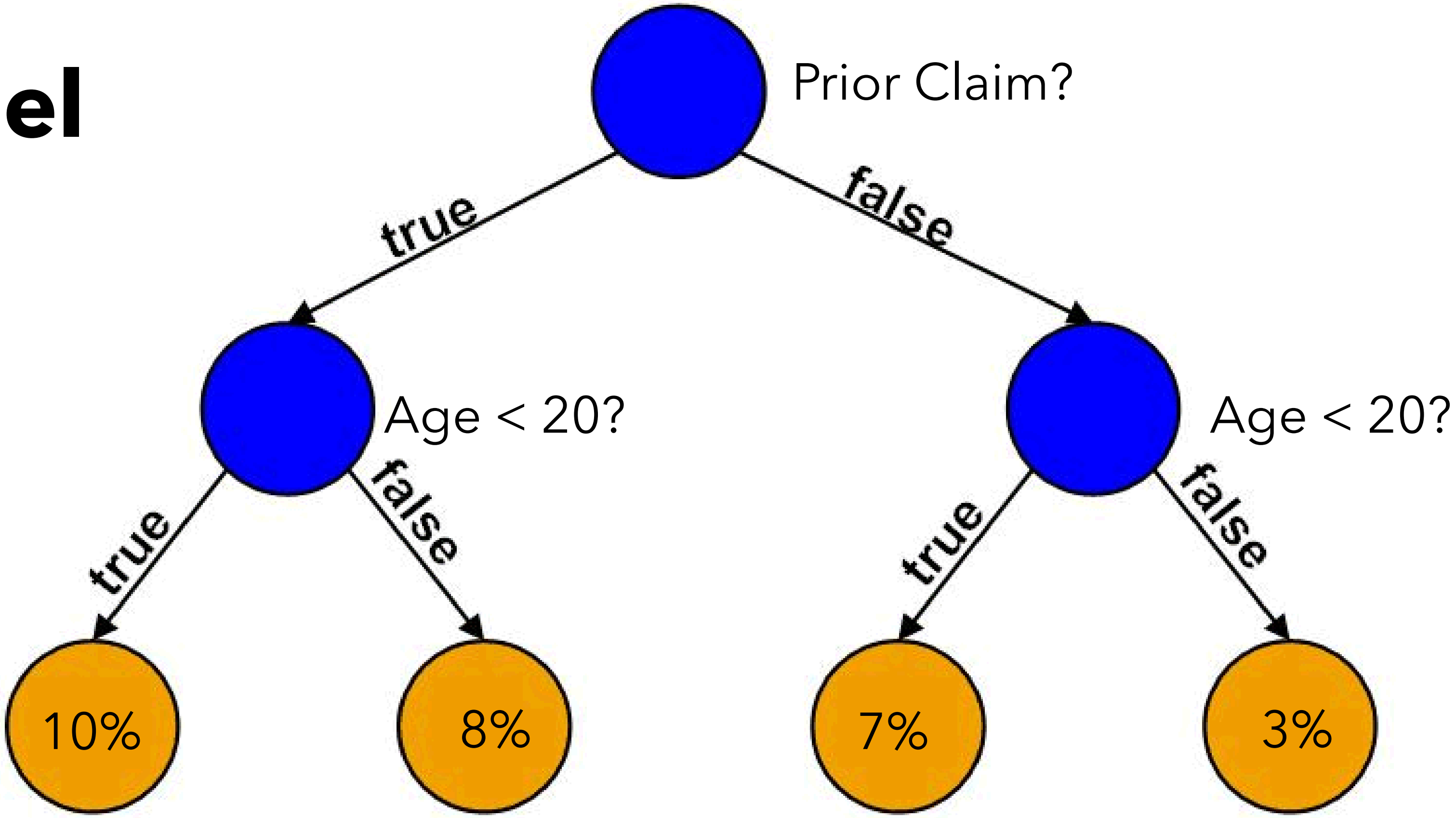
Tree Based Model

- Single Decision Tree
  - Easy to Understand
  - Mimics how people make decisions
  - Easily interpreted



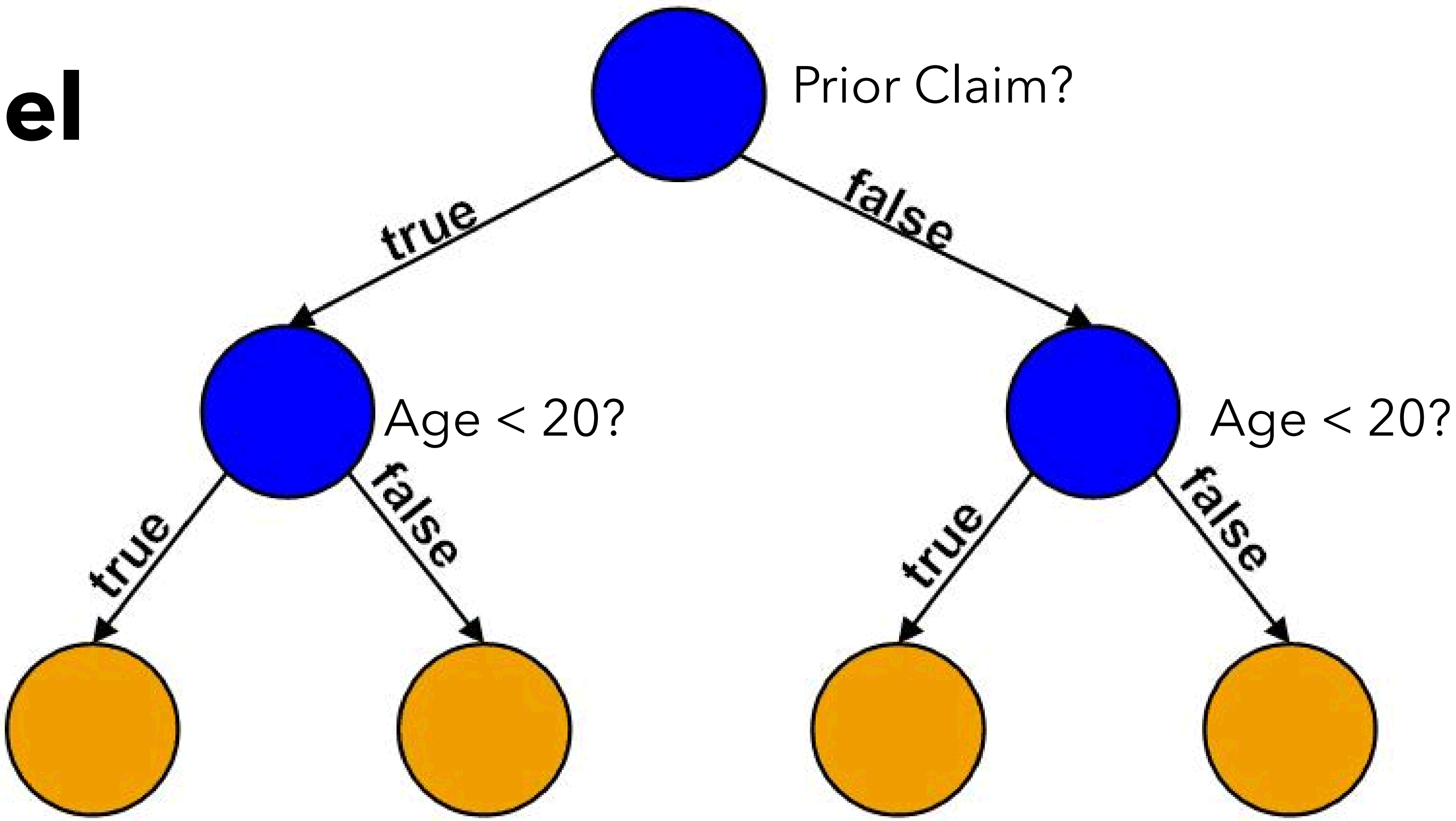
Tree Based Model

- Single Decision Tree
  - Easy to Understand
  - Mimics how people make decisions
  - Easily interpreted
- Classification returns a likelihood



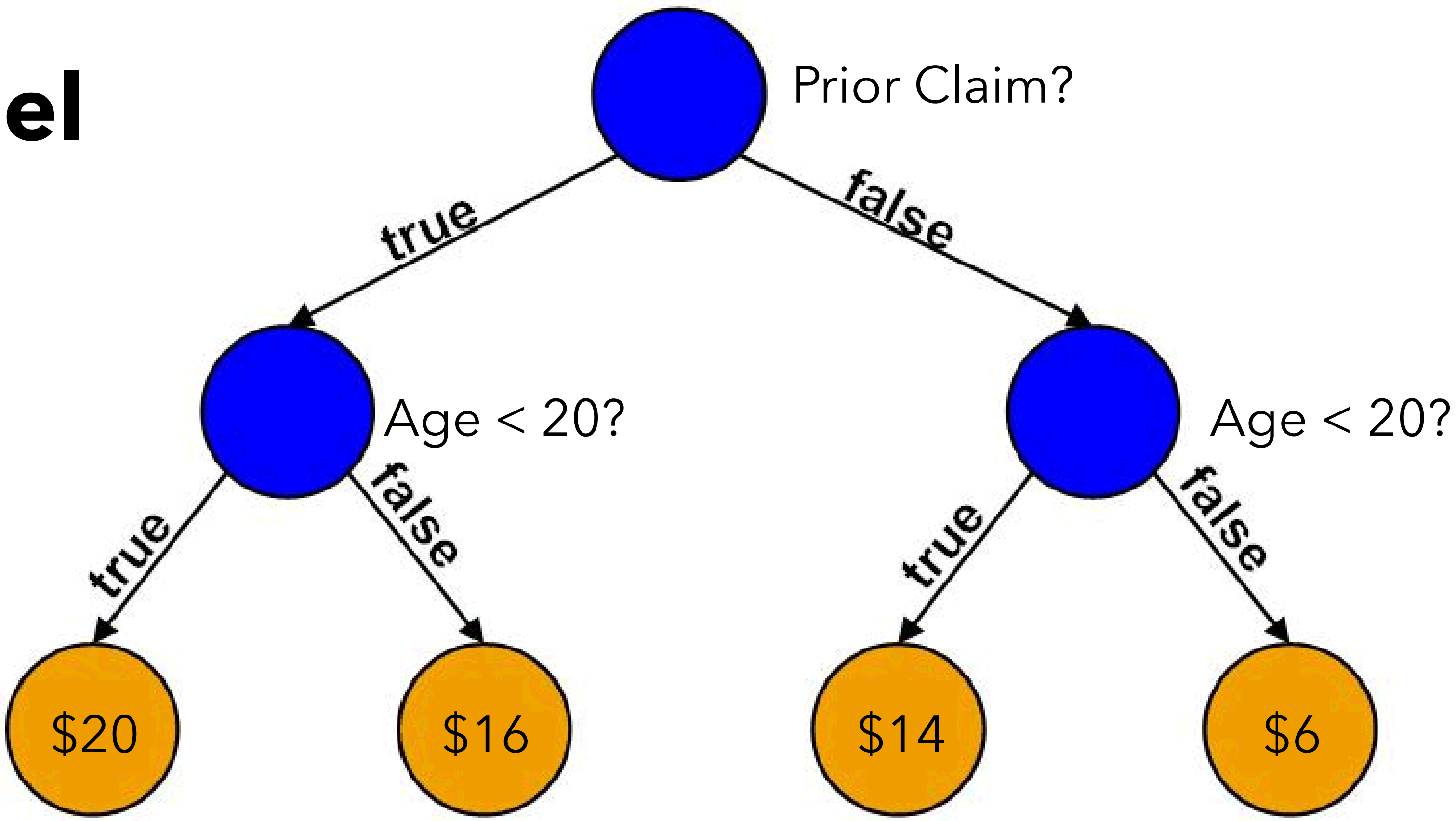
Tree Based Model

- Single Decision Tree
  - Easy to Understand
  - Mimics how people make decisions
  - Easily interpreted
- Classification returns a likelihood



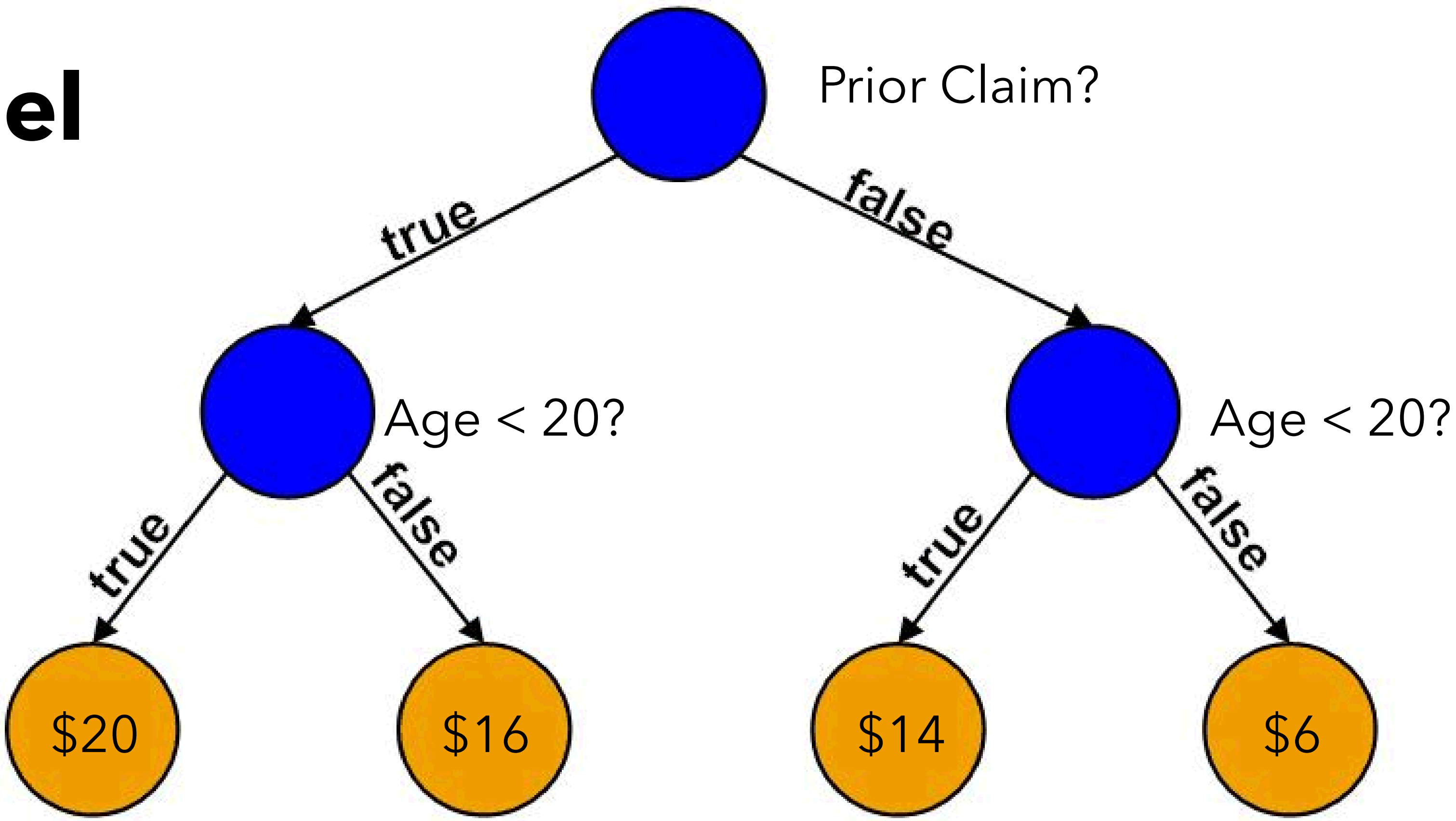
Tree Based Model

- Single Decision Tree
  - Easy to Understand
  - Mimics how people make decisions
  - Easily interpreted
- Classification returns a likelihood
- Regression returns a predicted amount



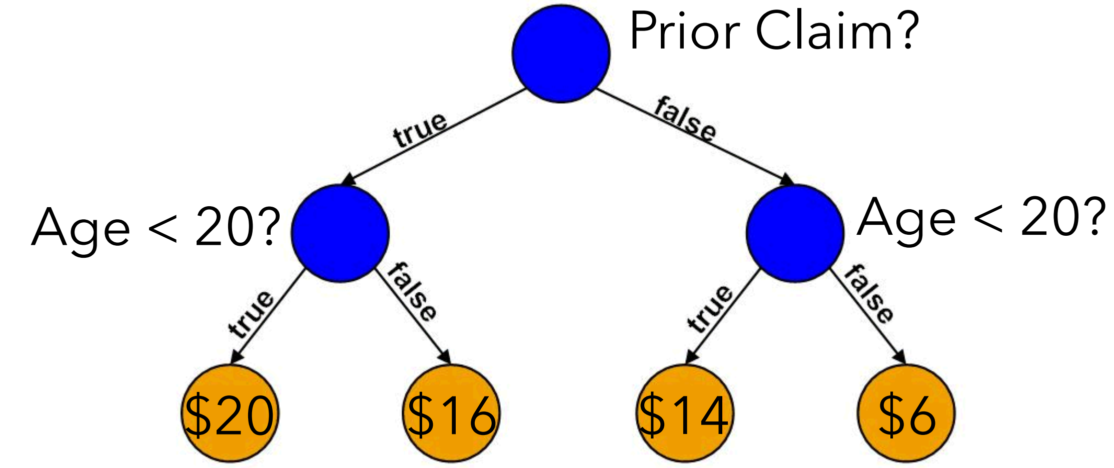
Tree Based Model

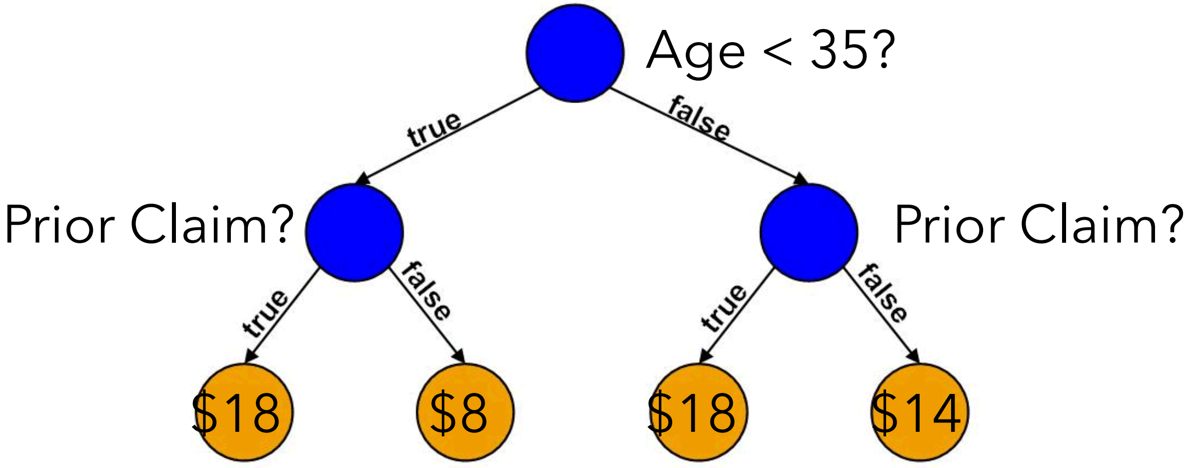
- Terminology
  - Nodes
    - Root
    - Sub-Node
    - Parent/Child
  - Splitting
    - Branch
    - Sub-Tree



Bagged Trees

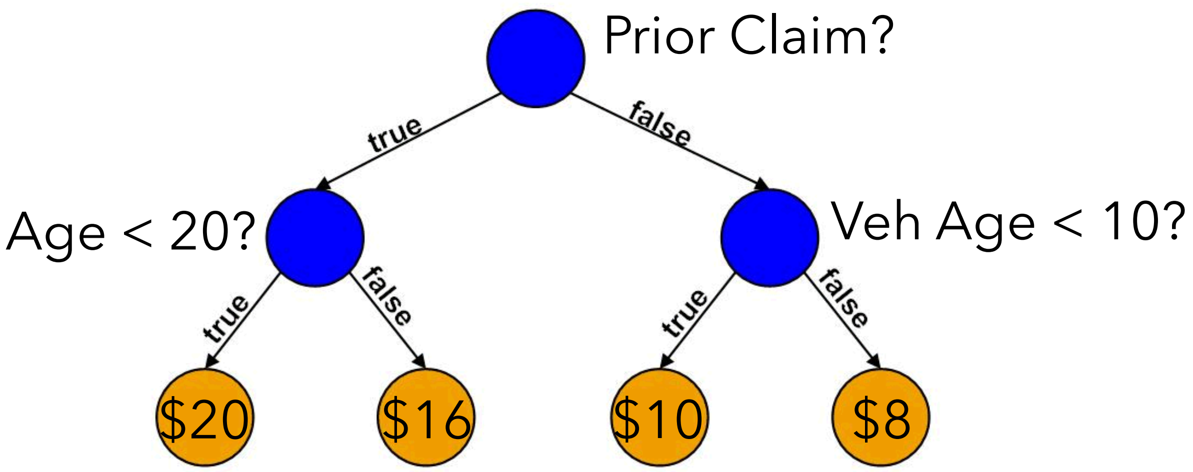
- Most Tree-Based Models are an "ensemble" of models
- "Bagged" Trees are based on multiple trees
  - "Bagged" comes from "bootstrap aggregated"
  - Each tree is grown the same way
  - The difference is each tree is based on a different bootstrap sample
  - The same variables are considered in each tree
  - Final prediction is the average of each tree's prediction

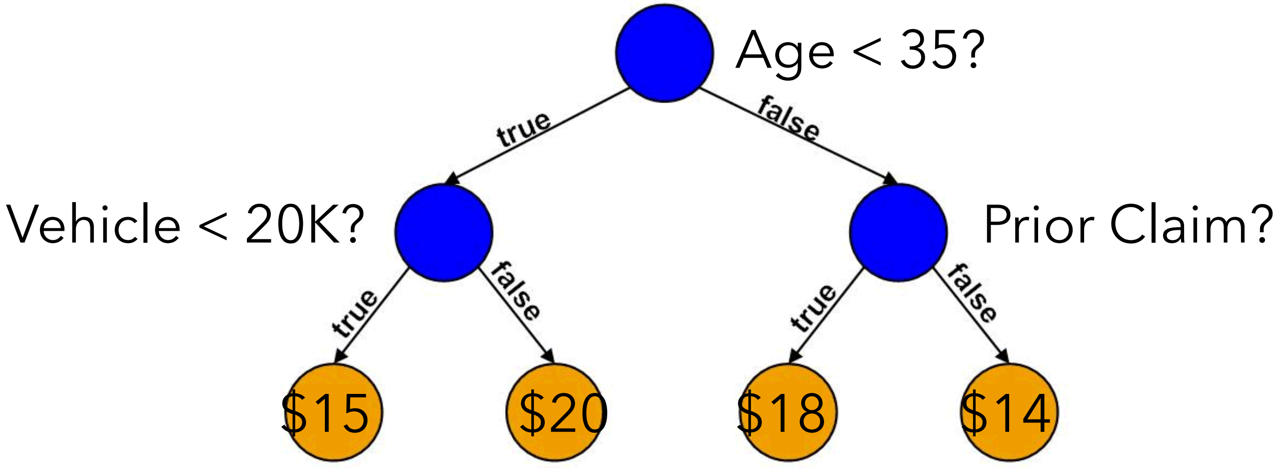




Random Forest

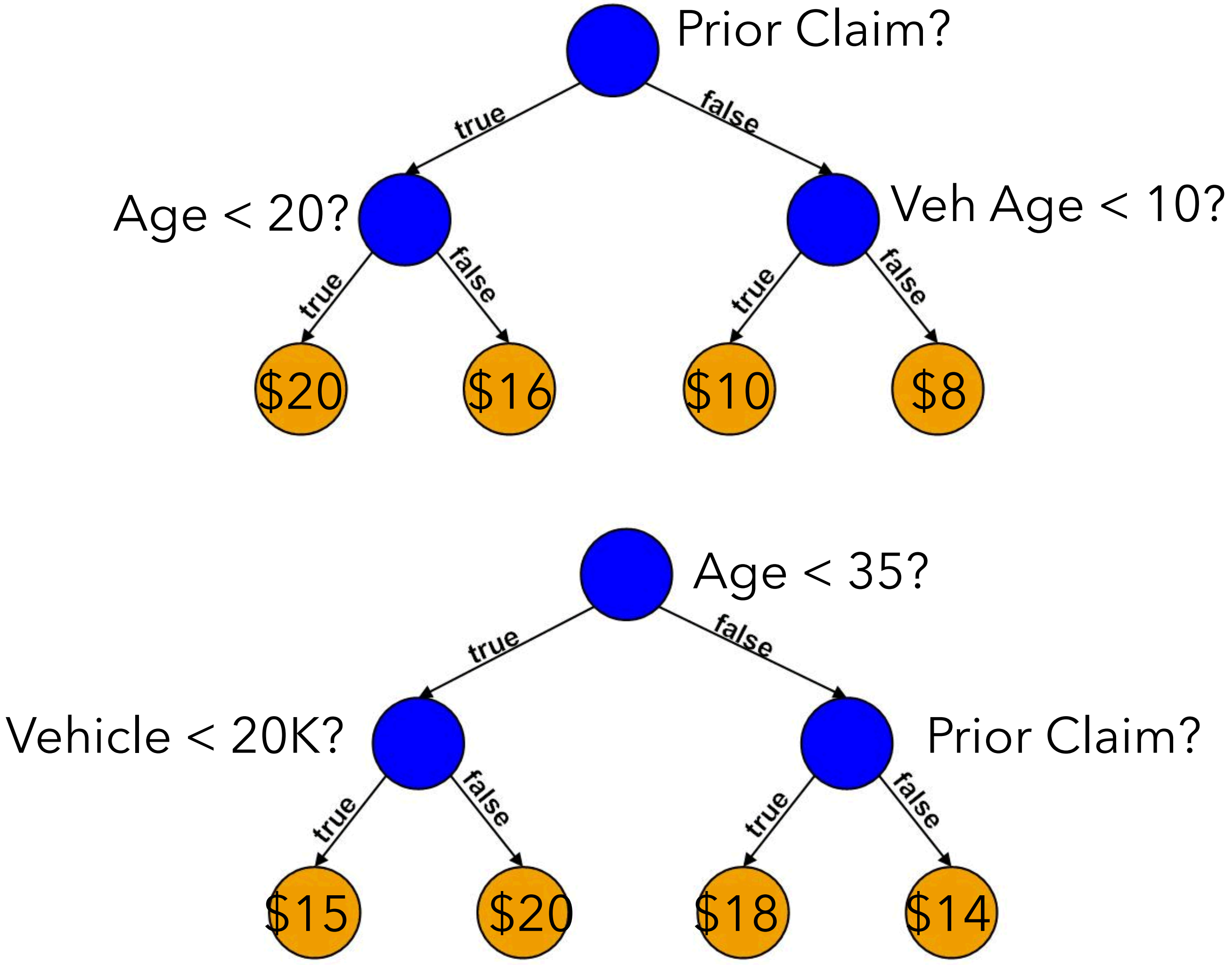
- Random Forest
  - Each tree is based on a different bootstrap sample (still)
  - Additionally: Randomly chosen variables considered at each split
  - Each tree is grown the same way
  - Final prediction is the average of each trees prediction
- Advantages
  - Trees are substantially different
    - Each tree not based on the same sample
    - Each split not based on the same variables





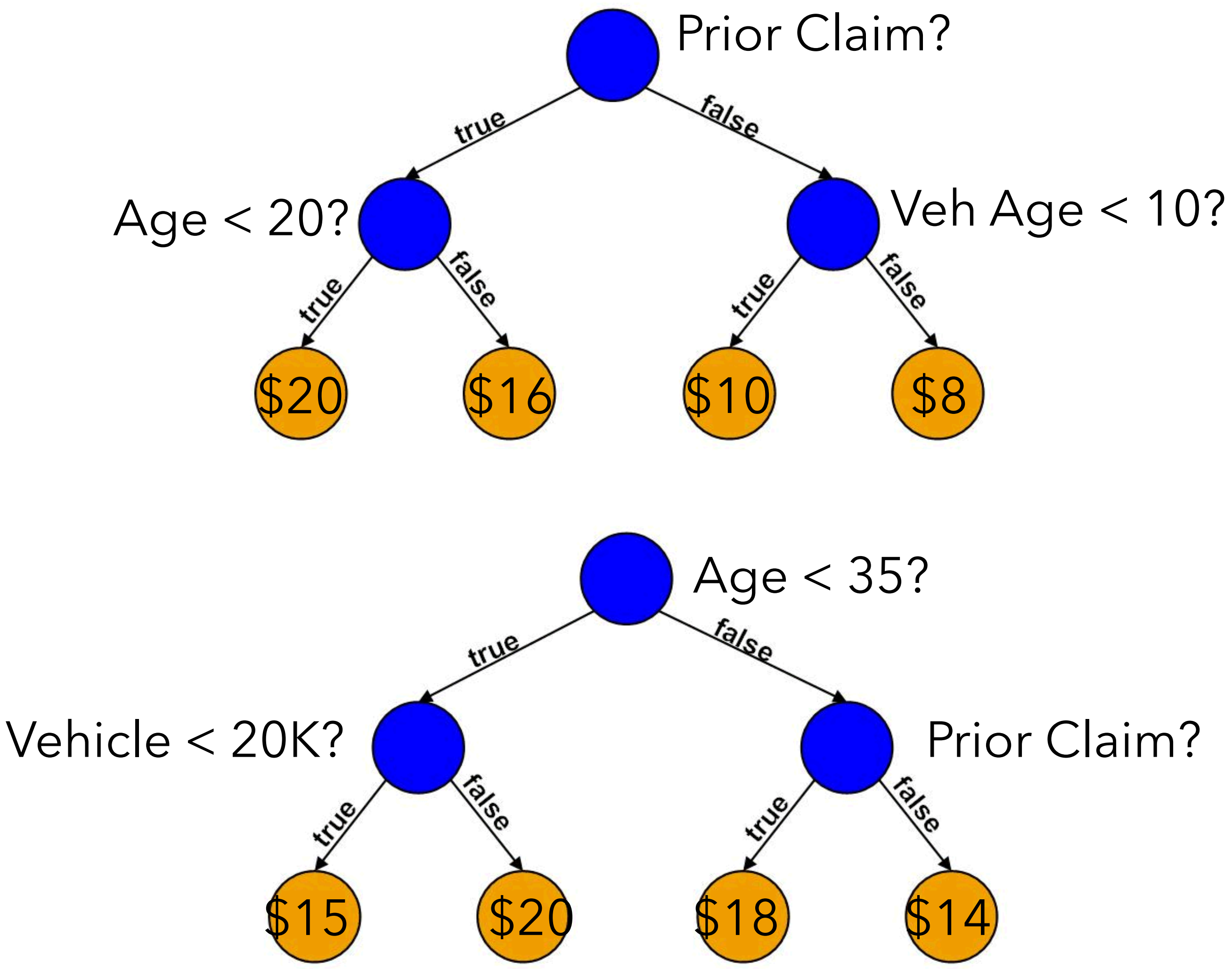
Random Forest

- Example
  - 22 year old driver, no prior claims
  - 5 year old vehicle, $15,000 vehicle
  - ($10+$15)/2 = $12.5



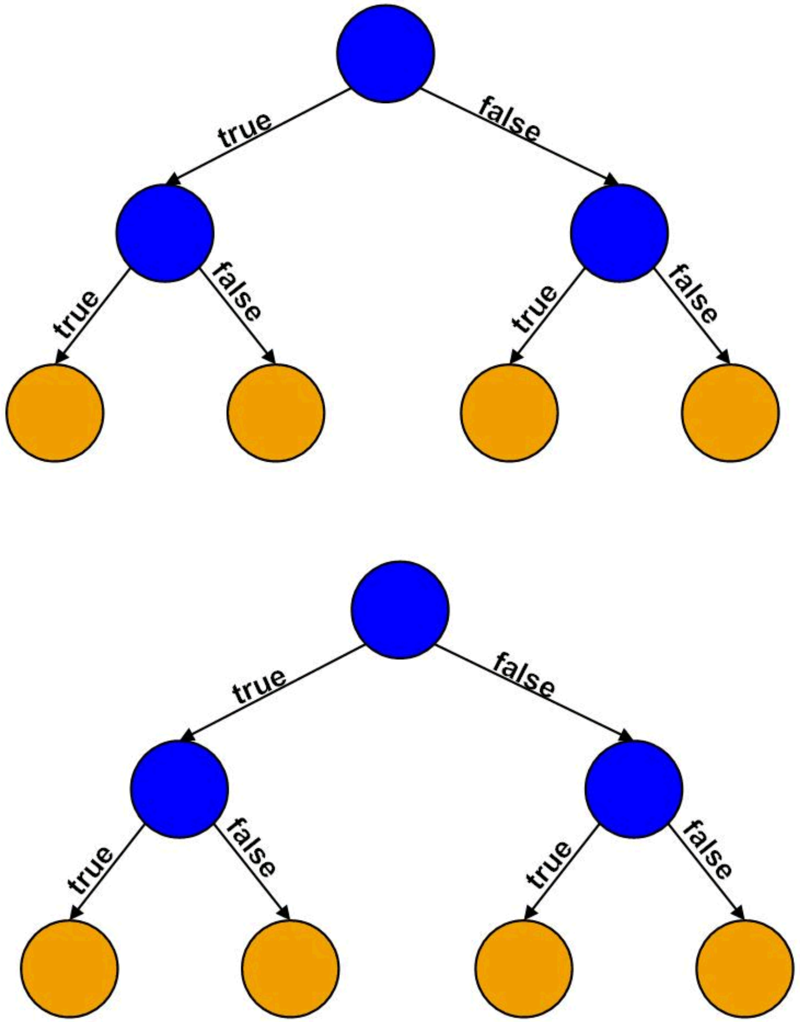
Random Forest

- Interpretation gets difficult
  - Trees can get very deep
  - There can be 100's or 1,000's of trees
- Many GLM statistical tests no longer apply
- There are many hyperparameters
  - Selections may materially impact the model
  - Selections should be checked for reasonability



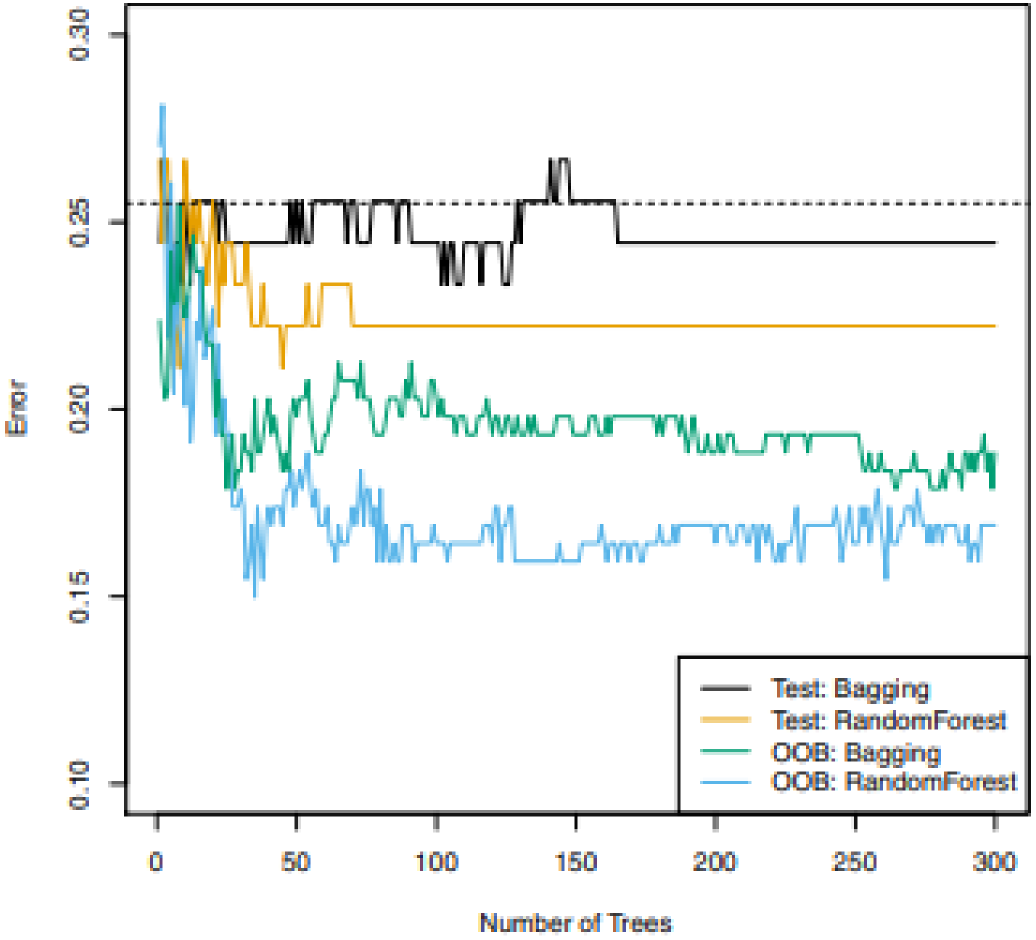
Random Forest

- Hyperparameters
  - Number of trees
  - Criteria on which to split
  - Bootstrap sample size (% of rows)
  - When to stop splitting
    - Max Tree Depth
    - Minimum Node Size
    - Max Leaf Nodes
  - Random Variables for each split (# of columns)



Random Forest

- Number of Trees
  - More trees makes the models more complex
  - The number of trees should be "tuned" to reduce error on either:
    - Separate test dataset
    - Out-Of-Bag data from training dataset
  - Different software may have different "early stopping" rules. Companies should be able to explain these rules.



Introduction to Statistical Learning: https://web.stanford.edu/~hastie/ISLRv2_website.pdf

Random Forest Challenges

- Interpretability
- Prone to Overfit
- Auditability

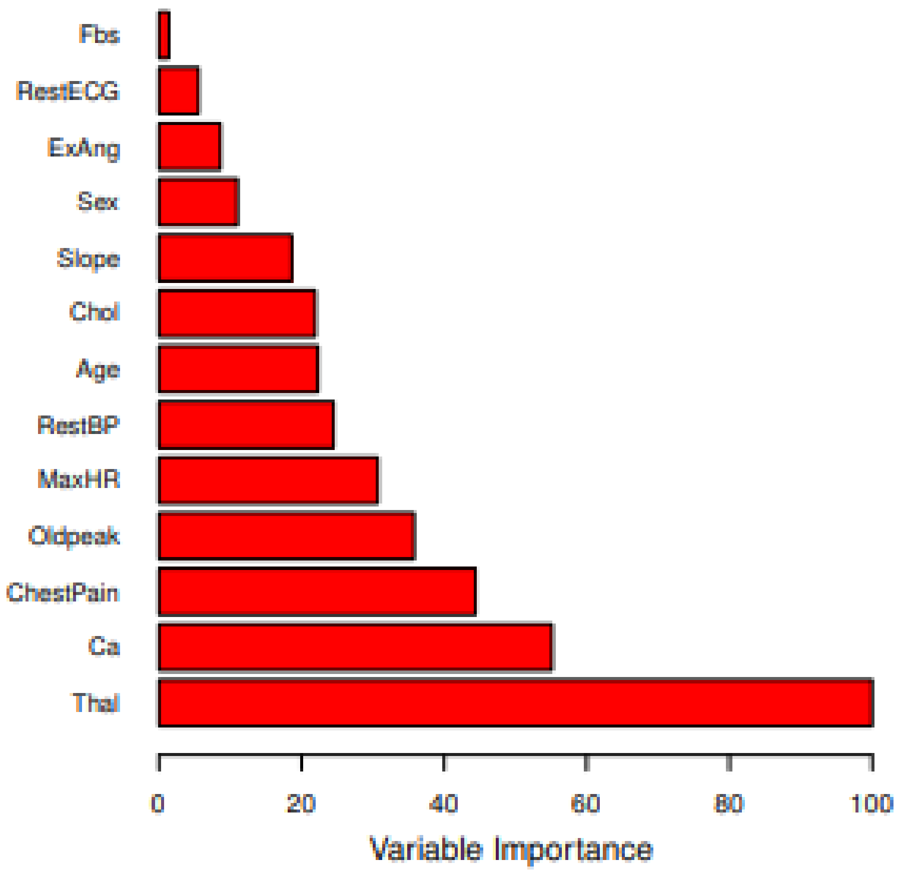
Challenges - Interpretability

- GLM's
  - Produce one set of model output
  - P-values provide a measure of statistical significance
    - Higher values can be prioritized for further review
  - Log-link model coefficients are easy to understand
    - Beta < 0 is a discount, Beta > 0 is surcharge

- RF's
  - It is hard to digest the net impact of a collection of trees
  - Variable Importance Plots highlight which variables are relatively less important
  - Interpretability plots help understand the impact of a variable upon the model

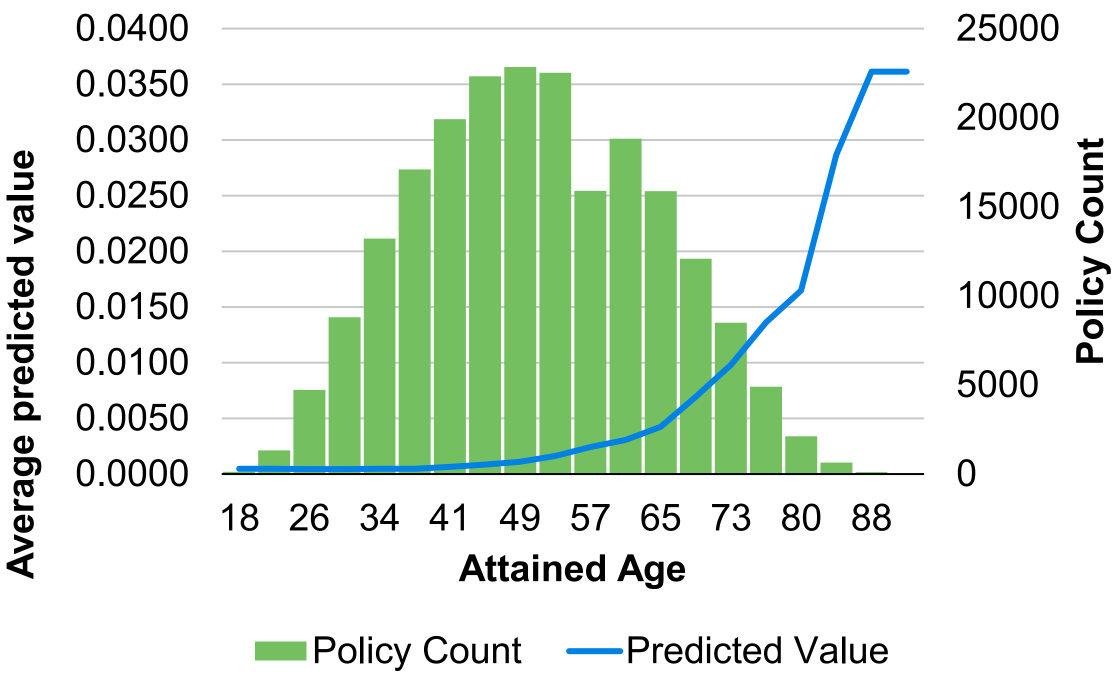
Variable Triaging

- Variable Importance Plots
  - Provide a measure of which variables are relatively more important than others
  - Variables with low importance measures aren't necessarily unimportant, but they might be
  - Further scrutiny may be appropriate for variables with a low importance measure
    - Similar to looking at variables with high p-values in a GLM
- Types of variable importance
  - Gain: improvement in prediction accuracy from feature
  - Cover: Number of observations influenced
  - Frequency: Number of times used to split data



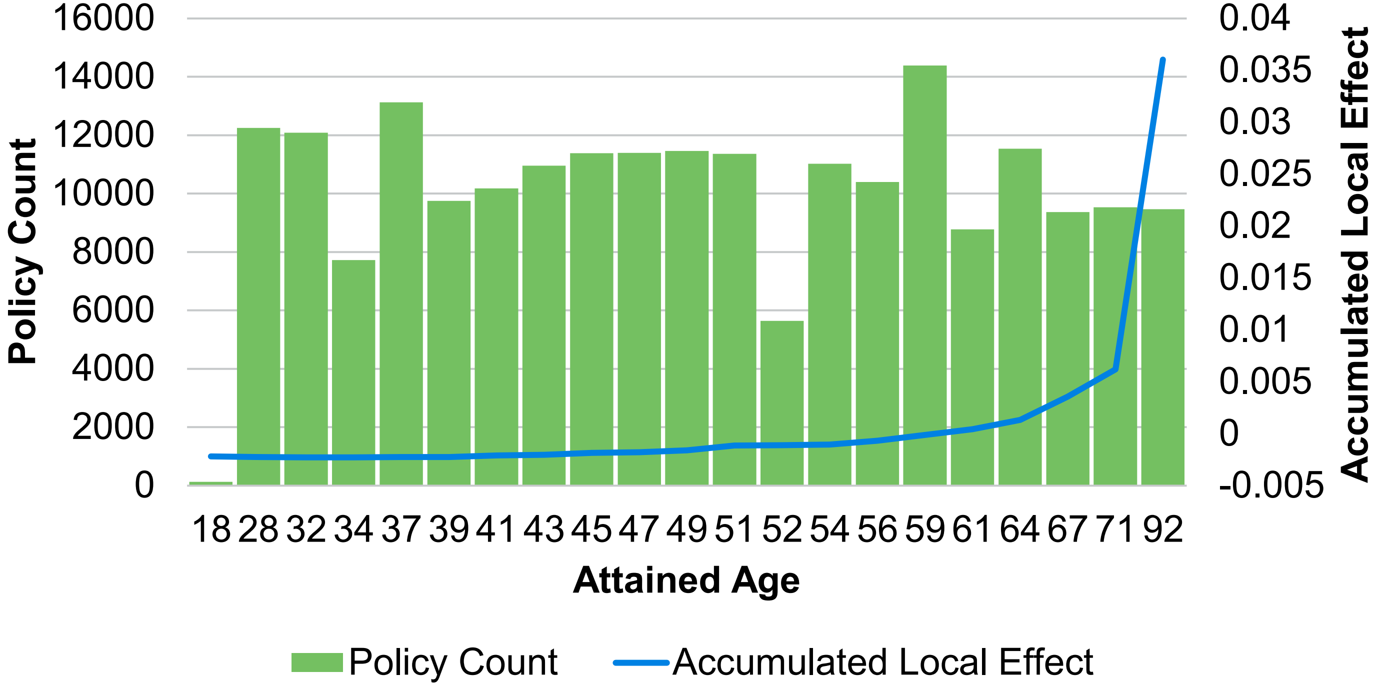
Interpretability Plots

- Partial Dependence Plots
  - Computes the marginal effect of a given feature on the prediction
  - Fixes the value of the predictor variable of interest, calculating the model prediction for each observation using the fixed value
  - Repeat for all values of the predictor variable



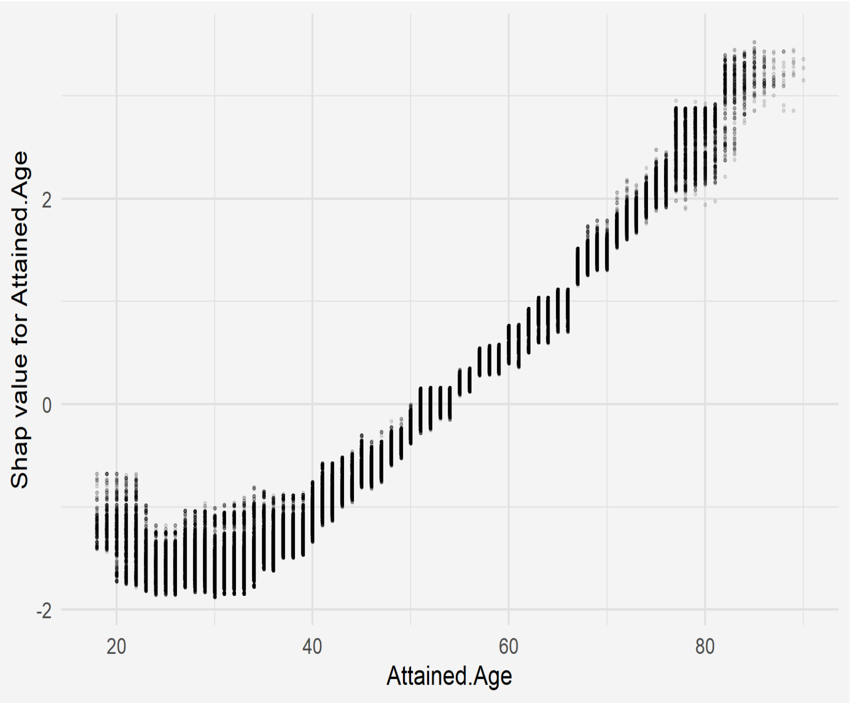
Interpretability Plots

- Accumulated Local Effects
  - Better option in the case of correlated features
  - Calculates and accumulates incremental changes in the feature effects
  - Shows the expected and centered effects of a feature, like a coefficient in a GLM



Interpretability Plots

- Shapely Additive Explanations
  - How much that feature moves the prediction away from the overall average prediction.



Feature increases predicted value higher than average value.

Feature decreases predicted value lower than average value.

Challenges – Prone to Overfit

- Review Hyperparameters
  - Number of trees should be large enough, but no larger
    - Look at plot to minimize OOB/Test Error or Deviance
  - Tree Complexity
    - Minimum node size should be set high enough for reasonable credibility
    - Rule of Thumb: Max depth of > 8 may be too high
  - Other hyperparameters should be disclosed and briefly commented on
    - Bootstrap sample size (% of rows)
    - Random Variables tried for each split (# of columns)
  - Criteria to split should match the model purpose (classification, regression)
- Review lift charts on test/holdout data

Challenges - Auditability

- GLM's
  - Indicated factors are reproducible if you have the coefficients and link function
  - Indicated factors can be stored in lookup tables
  - Auditing model predictions could easily be done, even for a large number of risks

- RF's
  - Complete documentation means diagrams or if statements representing every component tree
  - Sample calculations would include input variable values, each tree's result, and the final result (average of the component trees)
  - A full audit of the logic would likely involve a significant amount of coding

Challenges - Auditability

- Random Forest Documentation
  - Exhibits could be made for spot-checking against tree documentation
    - Input Predictors
    - Individual Tree Predictions
    - Overall Model Prediction (average)

Sample Risk	Driver Age	Prior Claims	Vehicle Age	…	Tree 1	Tree 2	Tree 3	…	Model Prediction
1	16	0	5	…	$ 50.00	$ 40.00	$ 30.00	…	$ 40.00
2	17	0	6	…	$ 49.00	$ 39.20	$ 29.40	…	$ 39.20
3	18	0	2	…	$ 48.02	$ 38.42	$ 28.81	…	$ 38.42
4	19	1	3	…	$ 47.06	$ 37.65	$ 28.23	…	$ 37.65
5	20	0	9	…	$ 46.12	$ 36.90	$ 27.67	…	$ 36.90

  - However, auditing every prediction for a book of business would still be extremely difficult

Draft Random Forest Appendix For Discussion

- Sending out 2 versions
  - Track Changes: Highlights removed, changed, and added items to the GLM Appendix
  - Final: Updated with the tracked changes for easy reading
- Looking for feedback for future Random Forest reviews

References

- Basic Decision Tree Terminology
  - https://medium.datadriveninvestor.com/the-basics-of-decision-trees-e5837cc2aba7
- Theoretical Introduction to Random Forest
  - Introduction to Statistical Learning (Chapter 8 – 8.2.2)
  - https://web.stanford.edu/~hastie/ISLRv2_website.pdf
- Interpretable Machine Learning (Variable Importance and Interpretability Plots)
  - https://us.milliman.com/-/media/milliman/pdfs/2021-articles/4-2-21-interpretable-machine-learning.ashx
  - Book Club Presentation: https://www.youtube.com/watch?v=-yMdTAIkewk
- Tree-Based Models Book Club: https://youtu.be/6UCbpAt4r9M