GLM vs. Machine Leaning

--- with Case Studies in Pricing

John(Jun) Zhou, Ph.D. FCAS CPCU

Debbie(Qianxin) Deng, FCAS

November 10-13, 2019, Honolulu, HI

华安保险
SINOSAFE INSURANCE

中国太平
CHINA TAIPING

西南財經大學
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

CAS

Antitrust Notice

- The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

- Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

- It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

Poll Questions

Poll 1: Have you ever done a GLM analysis in pricing?

- Yes
- No

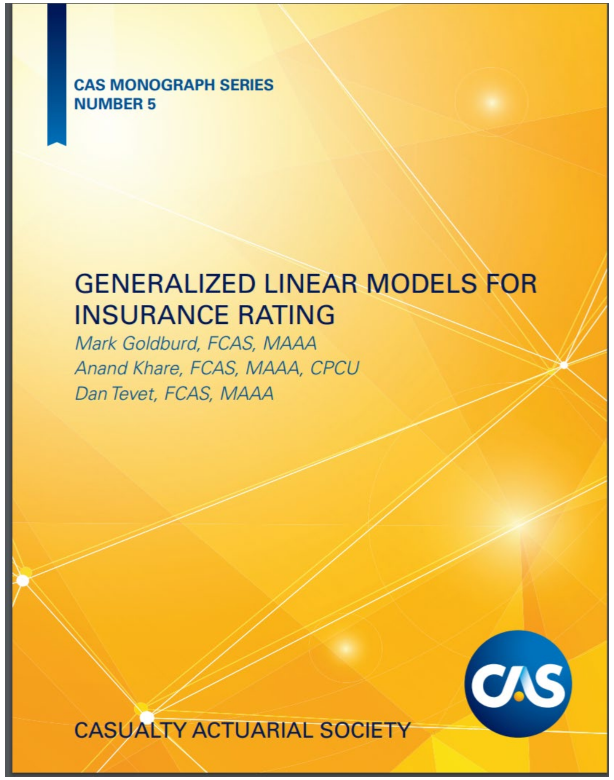
Poll 2: Have you ever done a Machine Learning analyses?

- Yes
- No

Contents

- A Quick Overview of GLM
- An Overview of Machine Learning
- Case Studies
- Summary

Generalized Linear Models



CAS MONOGRAPH SERIES
NUMBER 5

GENERALIZED LINEAR MODELS FOR INSURANCE RATING

Mark Goldburd, FCAS, MAAA
Anand Khare, FCAS, MAAA, CPCU
Dan Tevet, FCAS, MAAA

CASUALTY ACTUARIAL SOCIETY

With increases in computing power and access to big data, actuaries have in fact been using GLMs in the insurance rating process for many years.

The use of GLMs for classifying risks and rating personal lines business has increased tremendously in recent years and has spread to commercial lines business as well.

https://www.casact.org/pubs/monographs/papers/05-Goldburd-Khare-Tevet.pdf

A Quick Overview of GLM

- **Three components of GLM**
  - **Link Function:** a monotonic differentiable function
  - **Response variable Y:** has a distribution in exponential family
  - **Linear component:** $\boldsymbol{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \cdots}$

$$\boldsymbol{g(E[y]) = X\,\beta}$$

- **Key focus for modelers:**
  - To find the explanatory variable which has strong predictive power
  - To explain the model results with acceptable level of credibility

Pros and Cons of GLM in Pricing

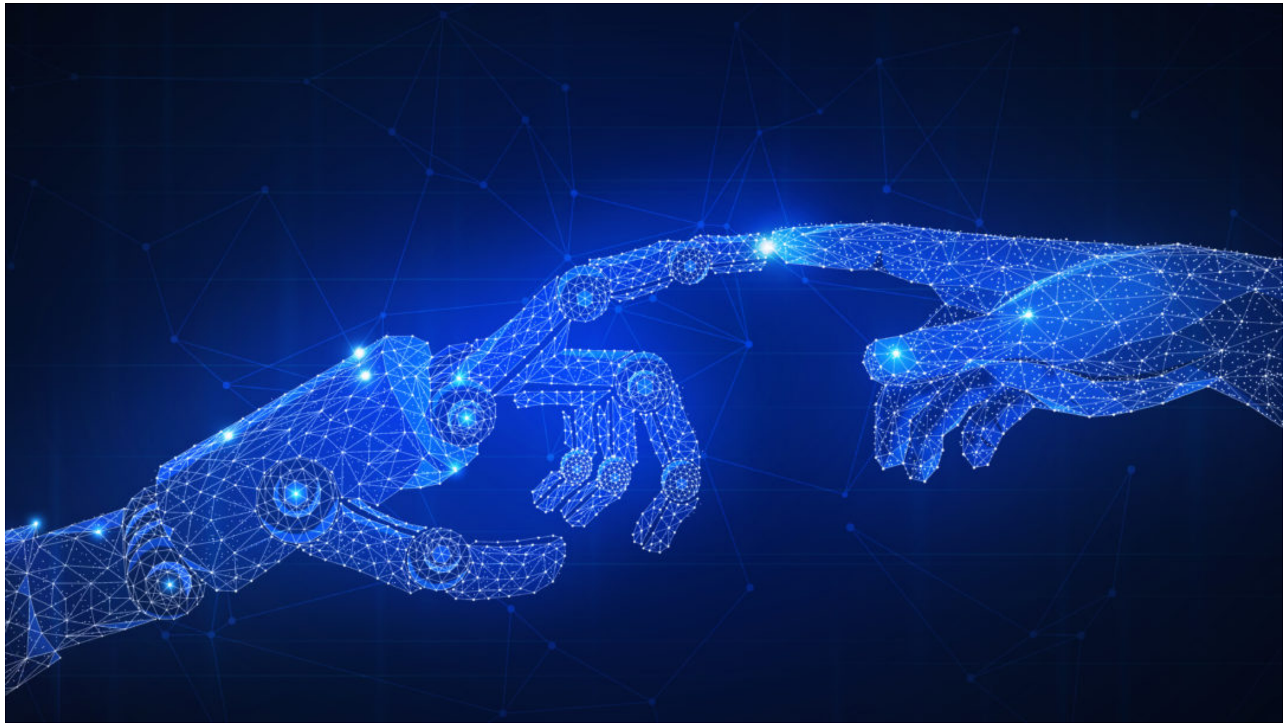
➢ **Pros**

- **Well established:** literature, regulatory acceptance, software, etc.
- **Empirically tested:** do find significant signals in insurance data
- **User-friendly:** adapt easily to rating manual and relativity concept

➢ **Cons of GLM:**

- **Assumptions:** assumptions, as link function, error function, underlying GLMs may not hold.
- **Interactions:** there is no systematic way to find all the relevant interactions.

Machine Learning



Machine learning is already all around us, unlocking our phones with a glance or a touch, suggesting music we like to listen to, and teaching cars to drive themselves, etc.

Artificial Intelligence (AI) has been described as the ‘fourth industrial revolution’.

https://www.information-age.com/artificial-intelligence-fourth-industrial-revolution-123475170/

What is Machine Learning?

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. **It is seen as a subset of artificial intelligence**.

Machine learning algorithms build a mathematical model based on sample data, known as "training data", **in order to make predictions or decisions** without being explicitly programmed to perform the task.

In its application across business problems, machine learning is also referred to as **predictive analytics**.

Reference:
https://en.wikipedia.org/wiki/Machine_learning

GLM vs. Machine Learning

- **Methodology:** Regarding prediction, GLM and machine learning can solve mostly the same problem from different perspectives.

- **Assumptions:** much less assumptions are needed for machine learning methods.

- **Predictability:** it is generally believed machine learning is superior than GLM.

A glance of ML algorithms

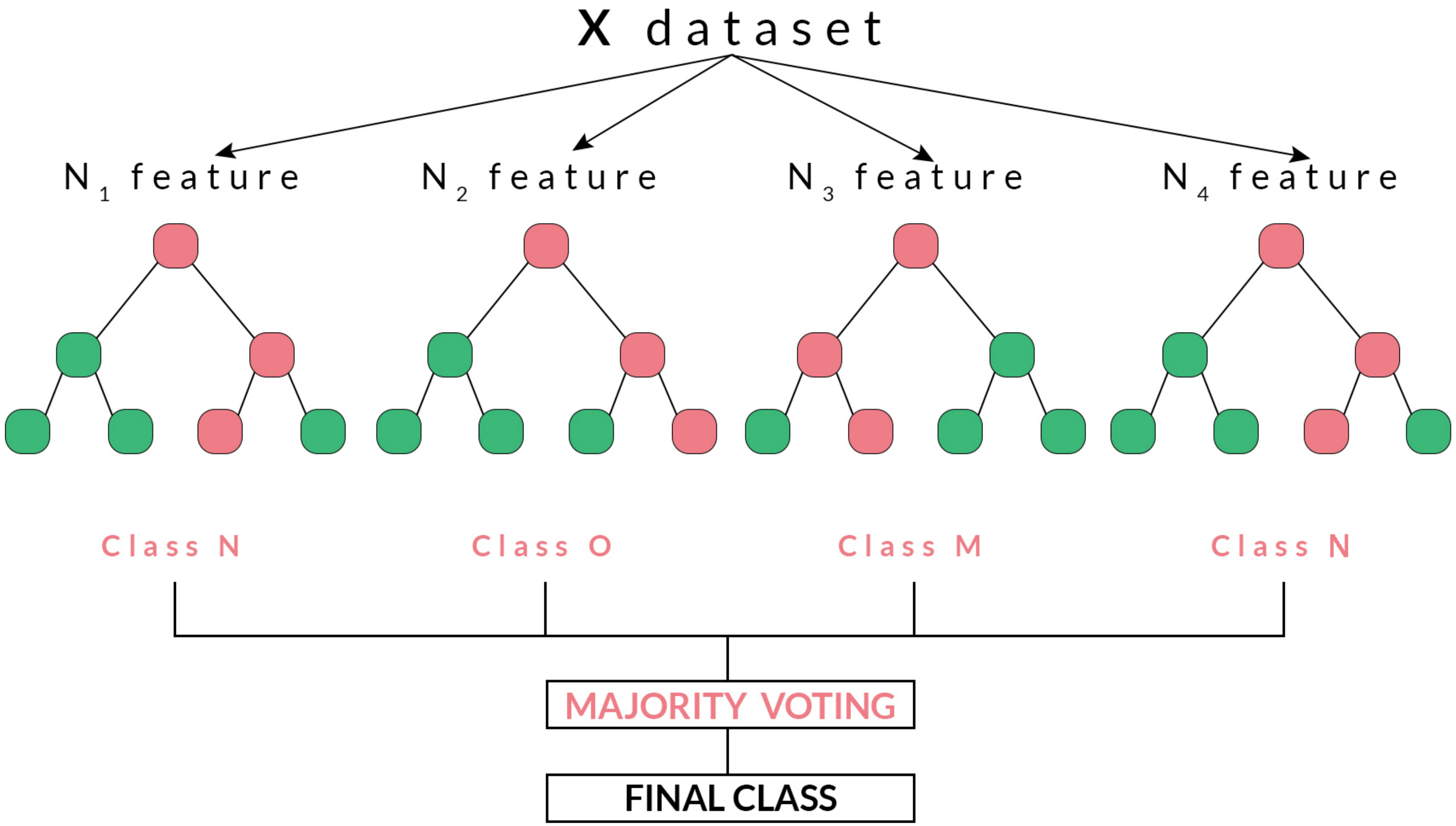
The types of machine learning algorithms differ **in their approach**, **the type of data** they input and output, and the **type of task or problem** that they are intended to solve.

Supervised learning algorithms： build a mathematical model of a set of data that contains **both the inputs and the desired outputs**.

Unsupervised learning algorithms： take a set of data that contains **only inputs**, and find structure in the data, like grouping or clustering of data points.

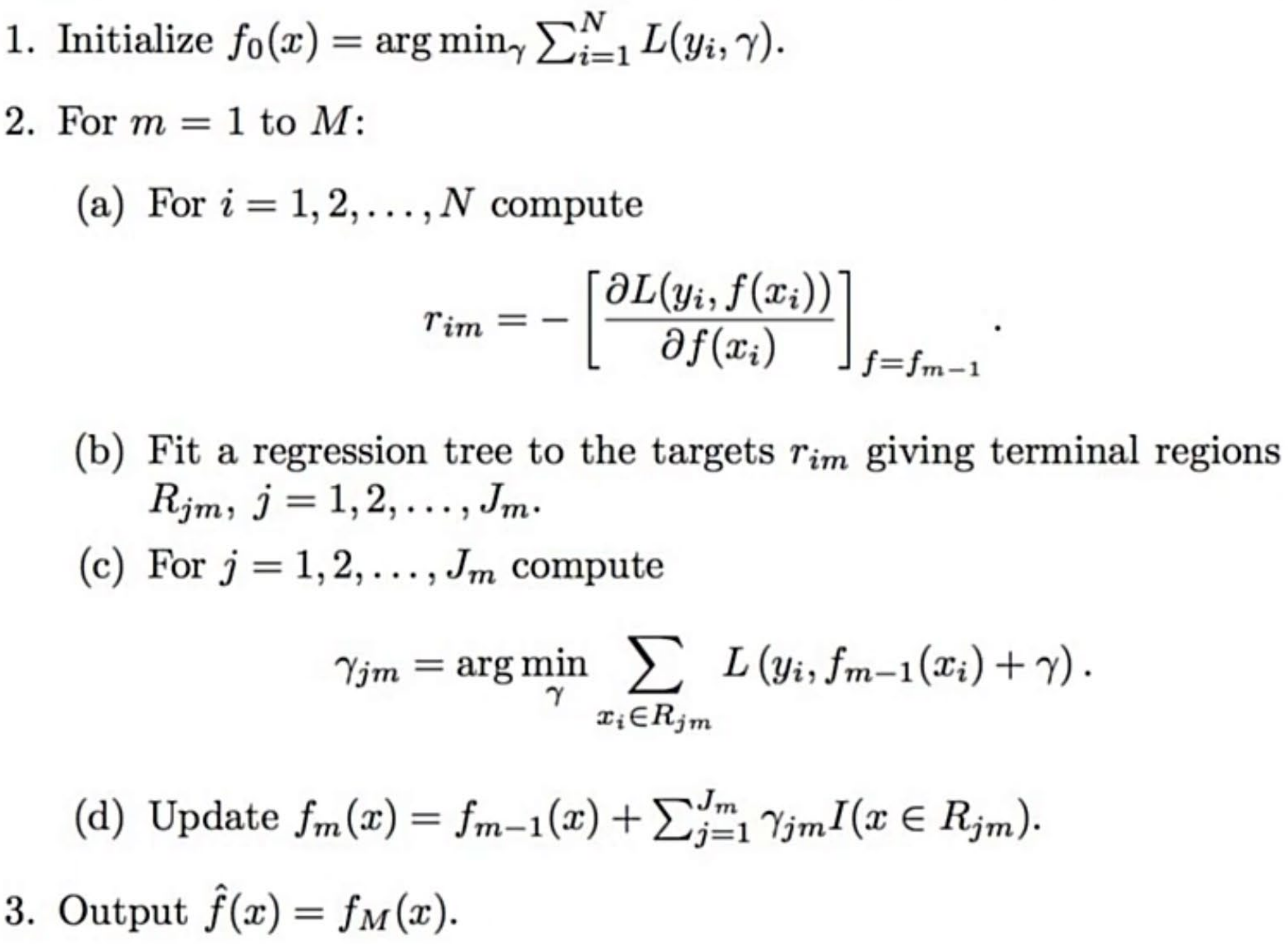
Reference:
https://en.wikipedia.org/wiki/Machine_learning

Random Forest (RF):

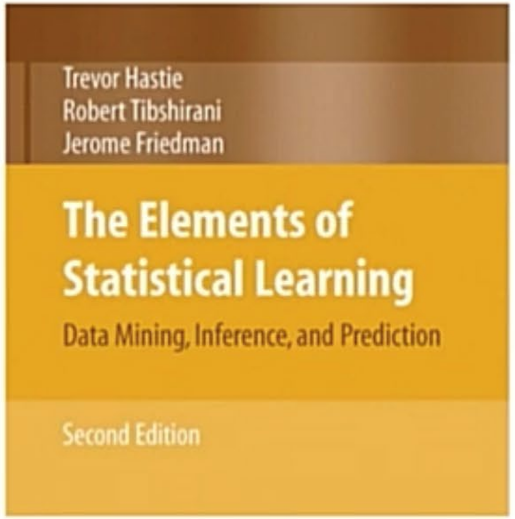


https://www.quantinsti.com/wp-content/uploads/2019/03/Random-Forest-Algorithm.jpg

Gradient Boosting Machine (GBM)



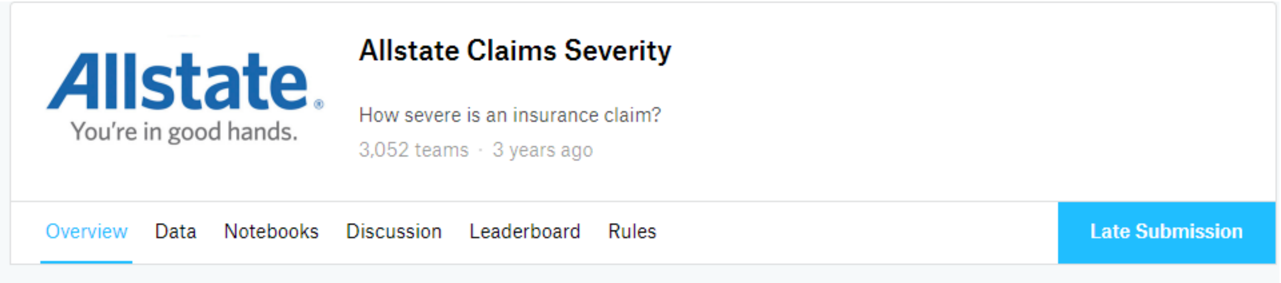
```
1. Initialize $f_0(x) = \arg\min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$.
2. For $m = 1$ to $M$:
    (a) For $i = 1, 2, \ldots, N$ compute
        $$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}.$$
    (b) Fit a regression tree to the targets $r_{im}$ giving terminal regions $R_{jm},\ j = 1, 2, \ldots, J_m$.
    (c) For $j = 1, 2, \ldots, J_m$ compute
        $$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L\left(y_i, f_{m-1}(x_i) + \gamma\right).$$
    (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.
3. Output $\hat{f}(x) = f_M(x)$.
```


Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition

Modelling Tools

- R packages
- Python scikit-learn
- H2O
- Xgboost
- Spark MLlib
- Vowpal Wabbit

Case Studies



Allstate
You're in good hands.

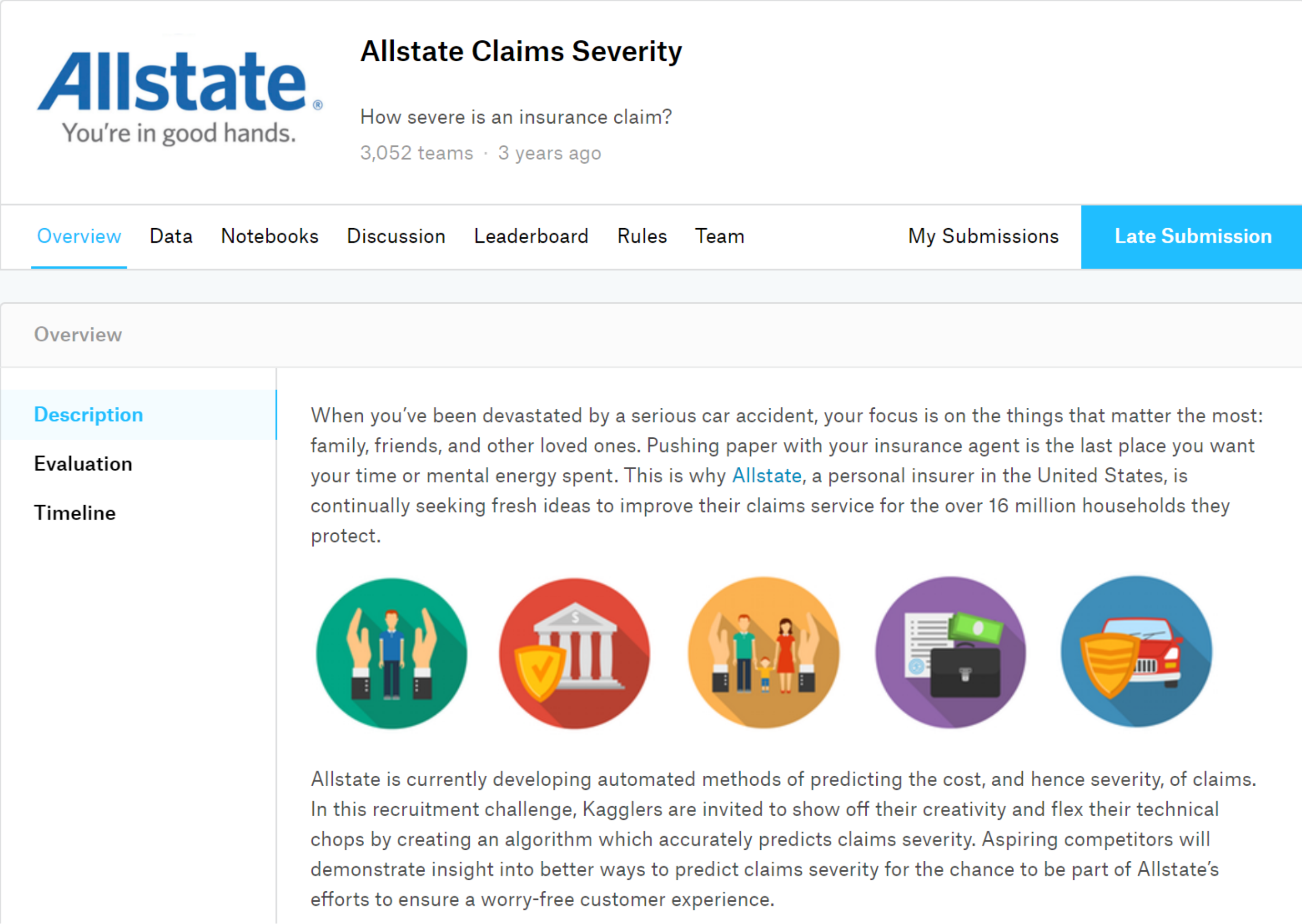
Allstate Claims Severity

How severe is an insurance claim?

3,052 teams · 3 years ago

2/3 of the winning solution in Kaggle competition use GBM

All State Claims Severity



Allstate Claims Severity

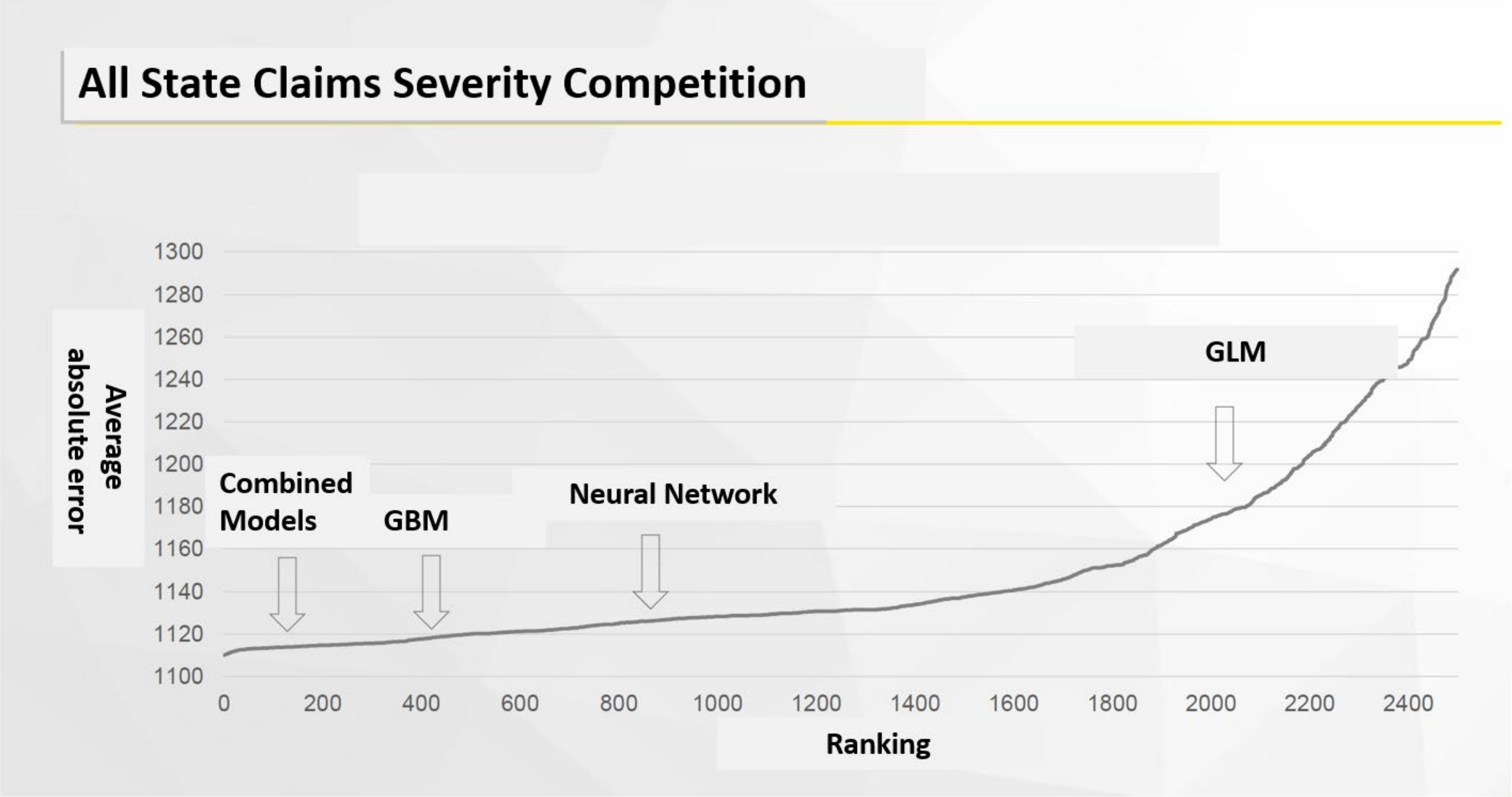
How severe is an insurance claim?

3,052 teams · 3 years ago

When you've been devastated by a serious car accident, your focus is on the things that matter the most: family, friends, and other loved ones. Pushing paper with your insurance agent is the last place you want your time or mental energy spent. This is why Allstate, a personal insurer in the United States, is continually seeking fresh ideas to improve their claims service for the over 16 million households they protect.

Allstate is currently developing automated methods of predicting the cost, and hence severity, of claims. In this recruitment challenge, Kagglers are invited to show off their creativity and flex their technical chops by creating an algorithm which accurately predicts claims severity. Aspiring competitors will demonstrate insight into better ways to predict claims severity for the chance to be part of Allstate's efforts to ensure a worry-free customer experience.

Kaggle Competition Ranking



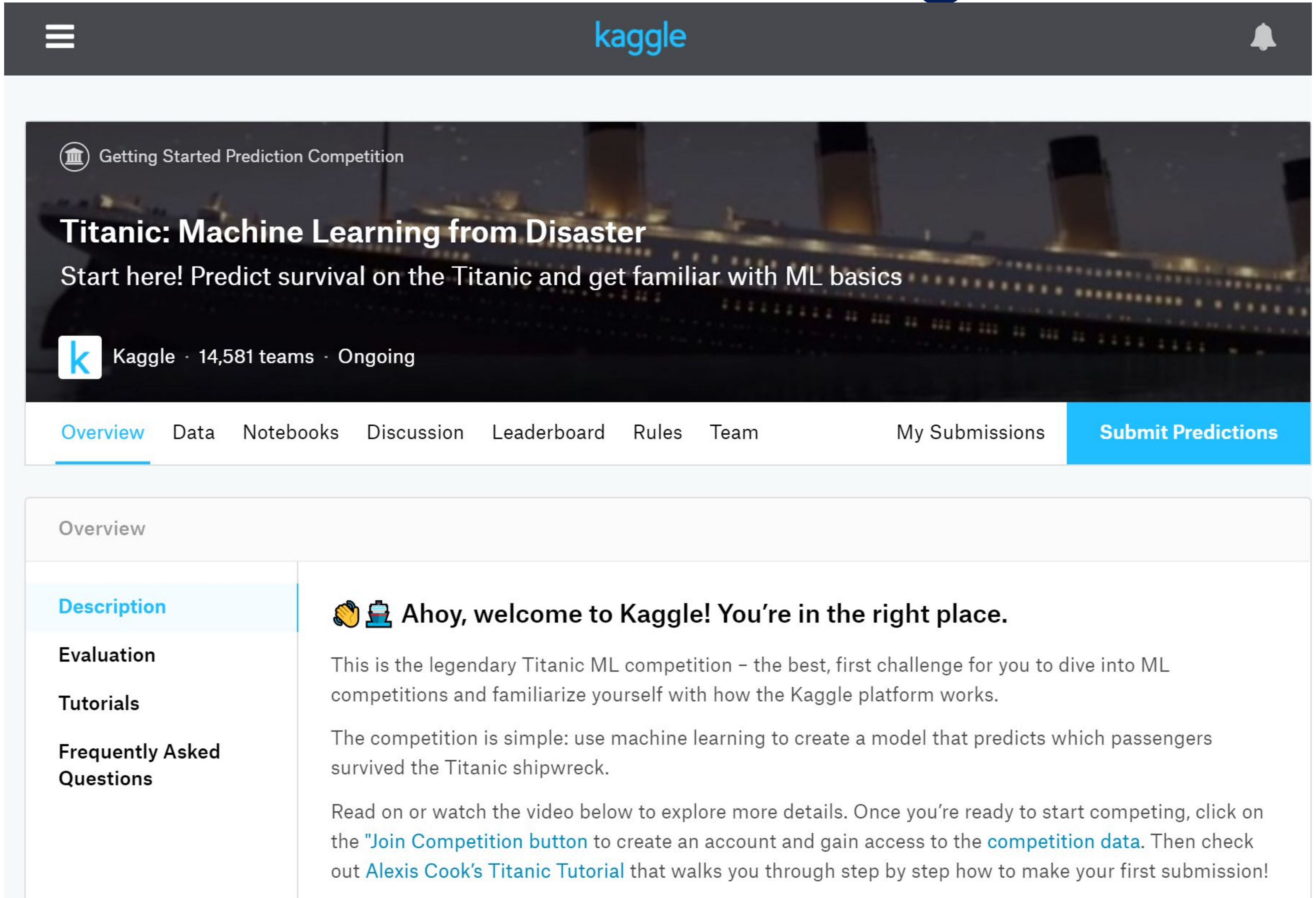
Reference:
Dr. Ji Yao’s unpublished research.

Winning Models

- #1st Place Solution:
  - w1*NN1^w2 + w3*NN2^w4 + w5*XGB1^w6 + w7 - weights optimized by using optim (Nelder-Mead) in a 1-fold manner => apply weights to test predictions => average 10 test predictions for 10x optimized weights.
  - If NN1 < w1 , then w2NN1^w3 + w4 Else if NN1 > w5, then w6NN1^w7+ w8 Else NN1
- #2nd Place Solution:
  - Level 1: The main ones were XGB and Keras NN (all of them with 4-6 bags)
  - Level 2: mainly trained XGB and Keras NN models, with different params, but also included linear regression with different target transformations, random forests and gradient boosting from sklearn
  - Level 3: quantile regression from statsmodels package
- #3rd Place Solution:
  - I ended up with using XGB and Keras exclusively for my final solution, which is an ensemble of around 100 base models (70% XGB & 30% Keras models). The test set predictions have been generated by a 20-times bagged Keras model with one hidden layer as stacker at the 2nd level.

https://www.kaggle.com/c/allstate-claims-severity/discussion

Start with Titanic Modeling



Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Kaggle · 14,581 teams · Ongoing

👋🚢 Ahoy, welcome to Kaggle! You're in the right place.

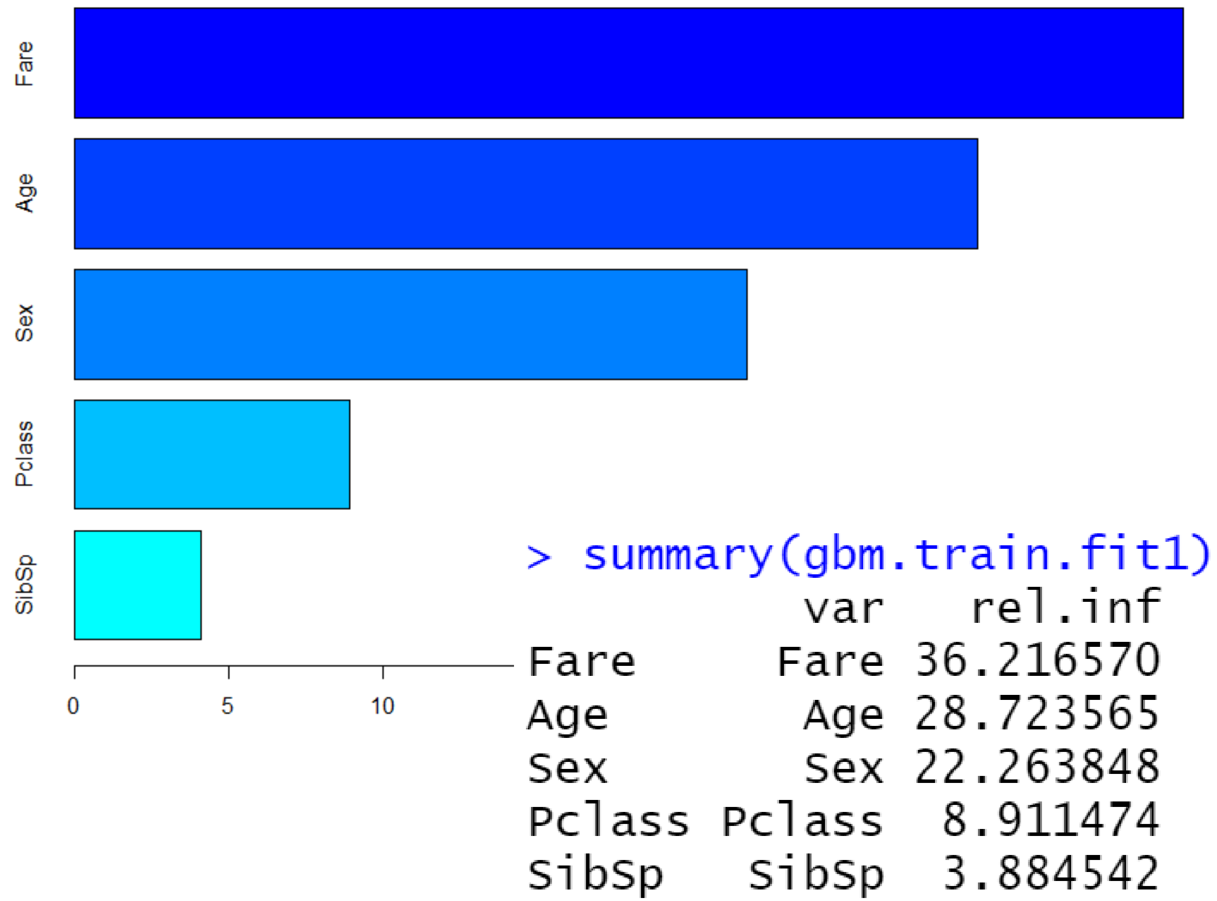
This is the legendary Titanic ML competition – the best, first challenge for you to dive into ML competitions and familiarize yourself with how the Kaggle platform works.

The competition is simple: use machine learning to create a model that predicts which passengers survived the Titanic shipwreck.

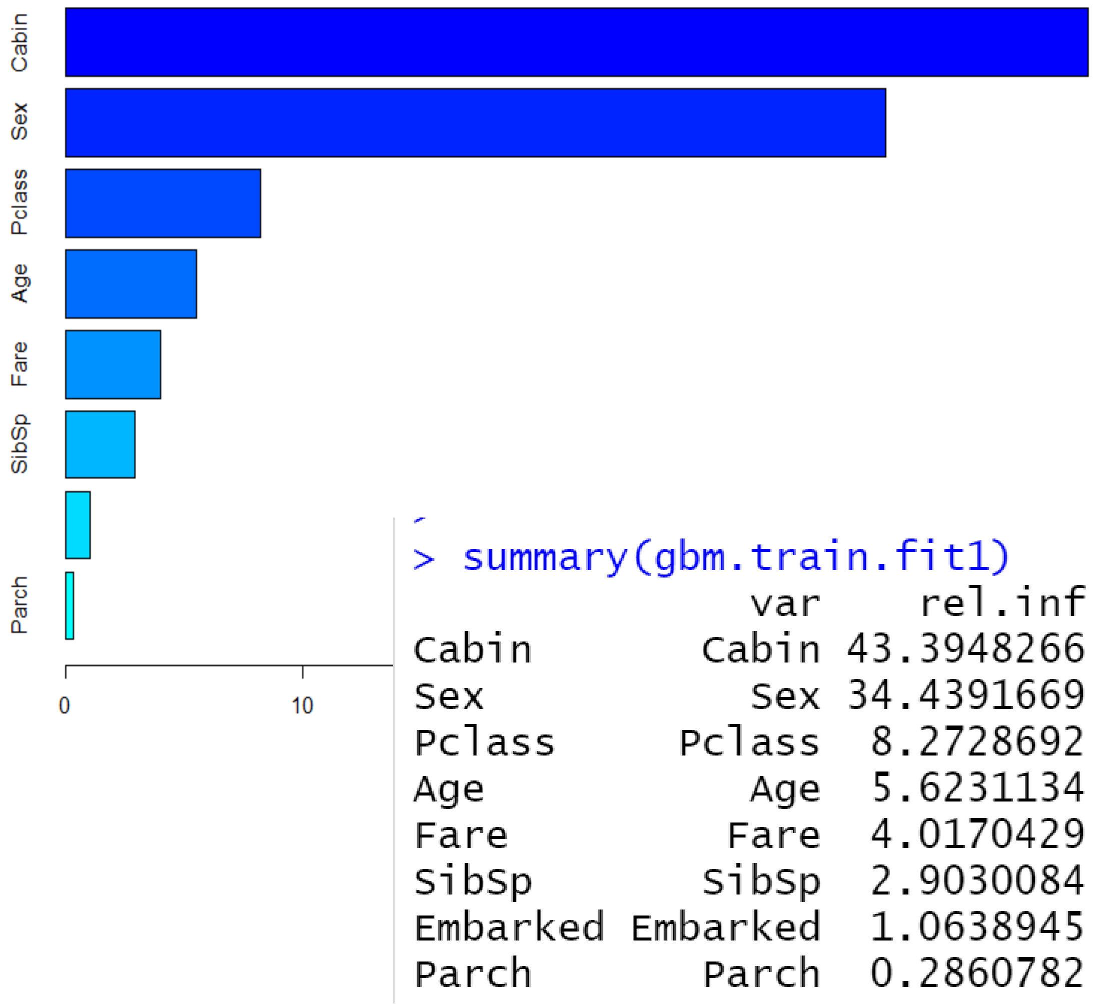
Read on or watch the video below to explore more details. Once you're ready to start competing, click on the "Join Competition button to create an account and gain access to the competition data. Then check out Alexis Cook's Titanic Tutorial that walks you through step by step how to make your first submission!

https://www.kaggle.com/c/titanic/data

Influence of variables

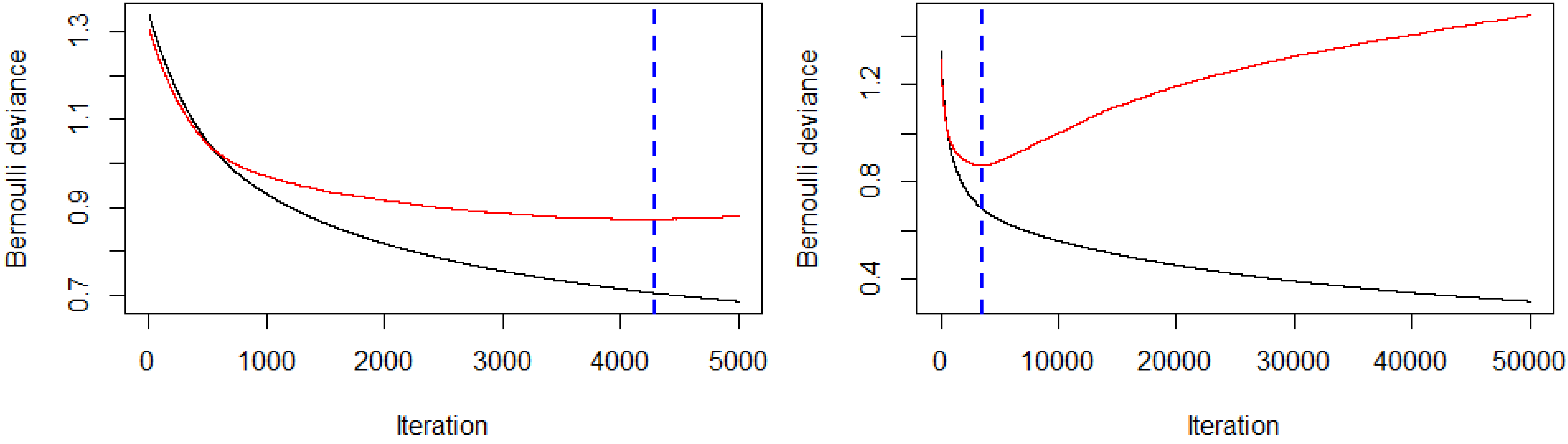


```
> summary(gbm.train.fit1)
          var   rel.inf
Fare     Fare 36.216570
Age       Age 28.723565
Sex       Sex 22.263848
Pclass Pclass  8.911474
SibSp   SibSp  3.884542
```



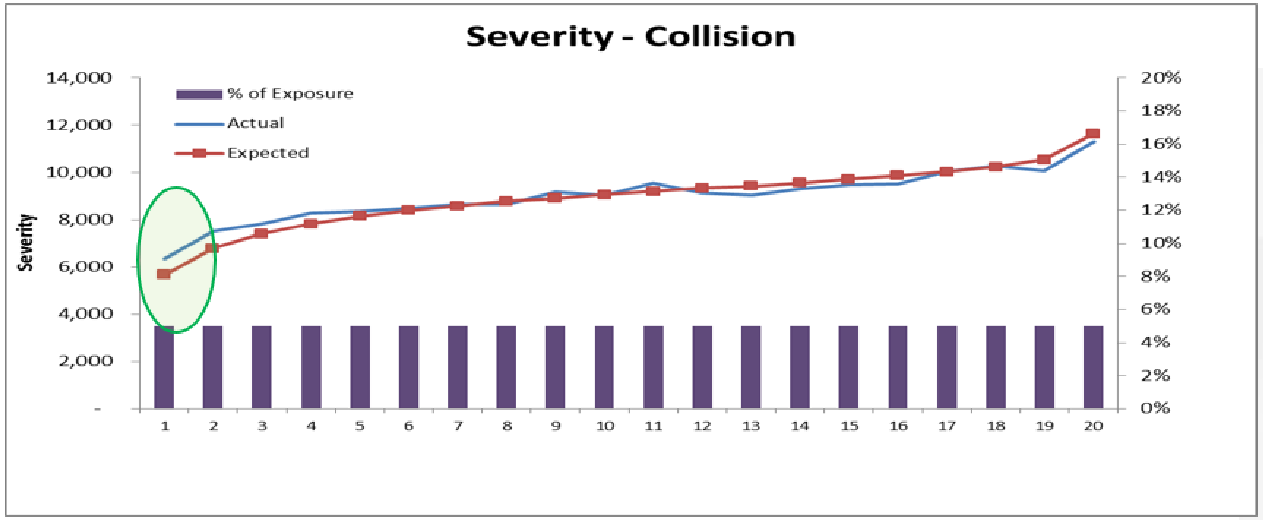
```
> summary(gbm.train.fit1)
              var    rel.inf
Cabin       Cabin 43.3948266
Sex           Sex 34.4391669
Pclass     Pclass  8.2728692
Age           Age  5.6231134
Fare         Fare  4.0170429
SibSp       SibSp  2.9030084
Embarked Embarked  1.0638945
Parch       Parch  0.2860782
```

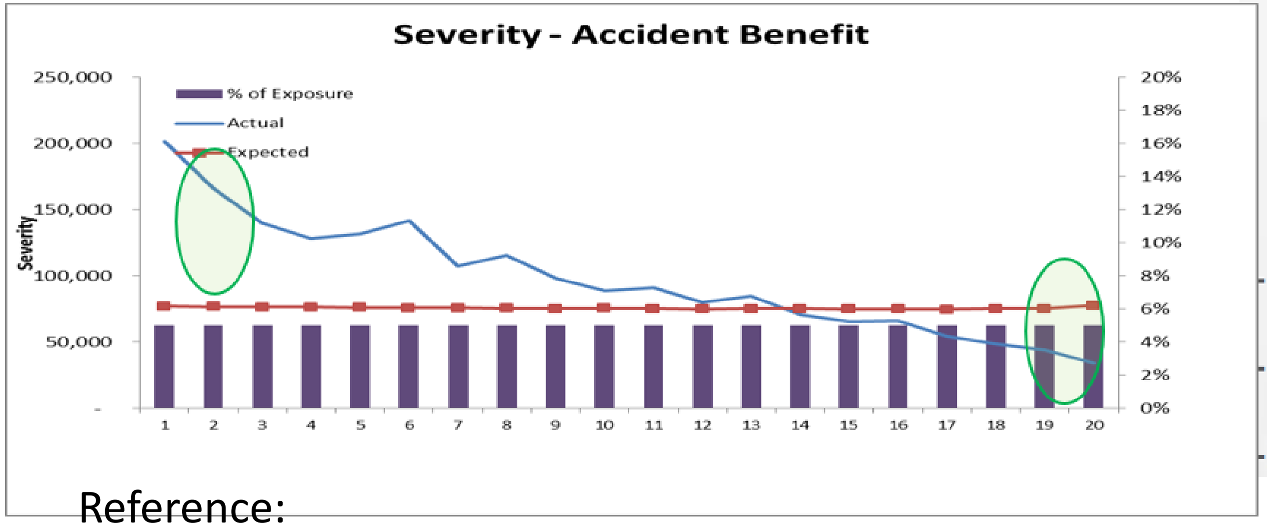
Overfitting --- Number of Iterations

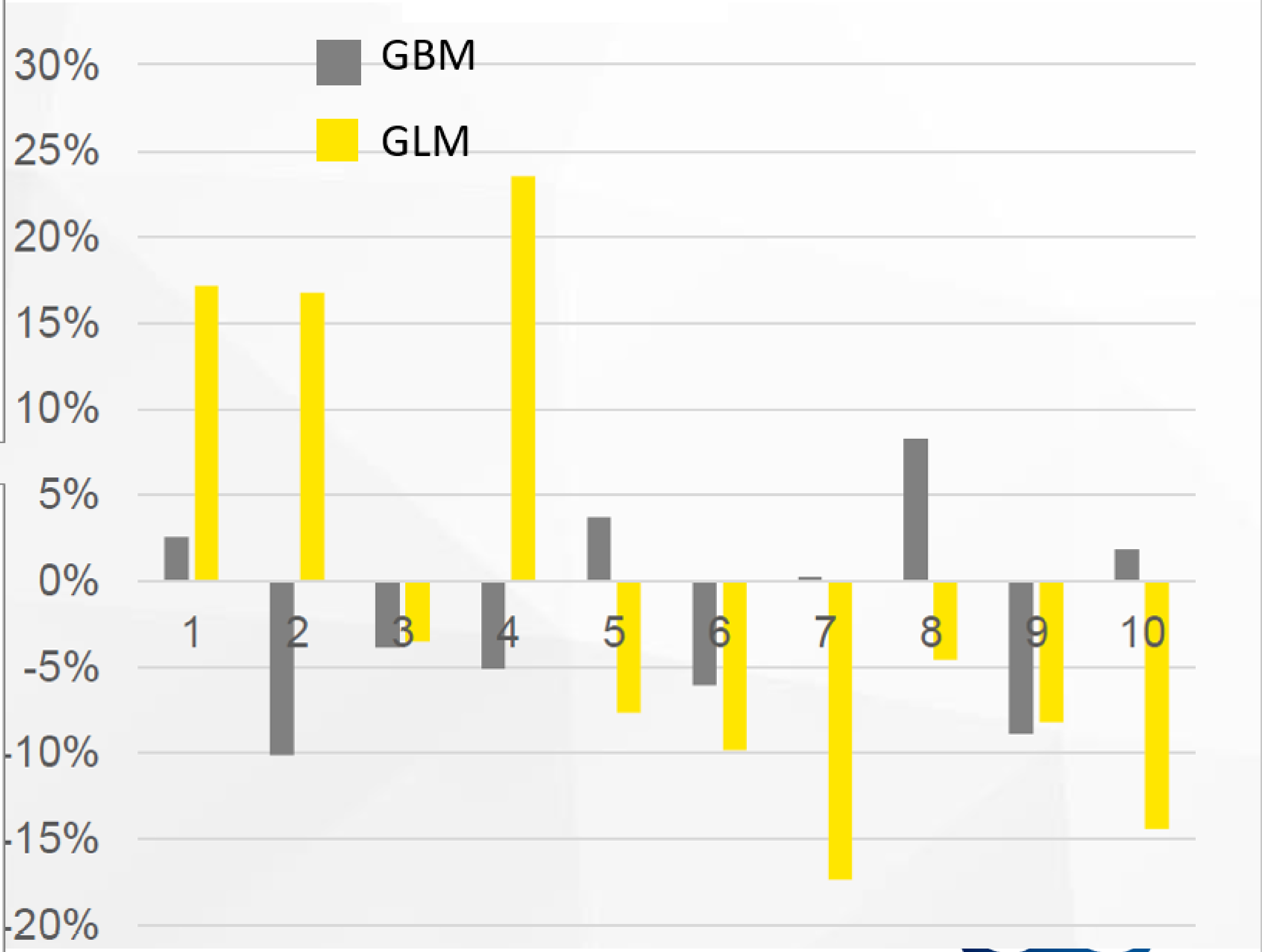


gbm.perf(object, plot.it = TRUE, oobag.curve = FALSE, overlay = TRUE, method)

Lift Graphs --- GLM vs. GBM







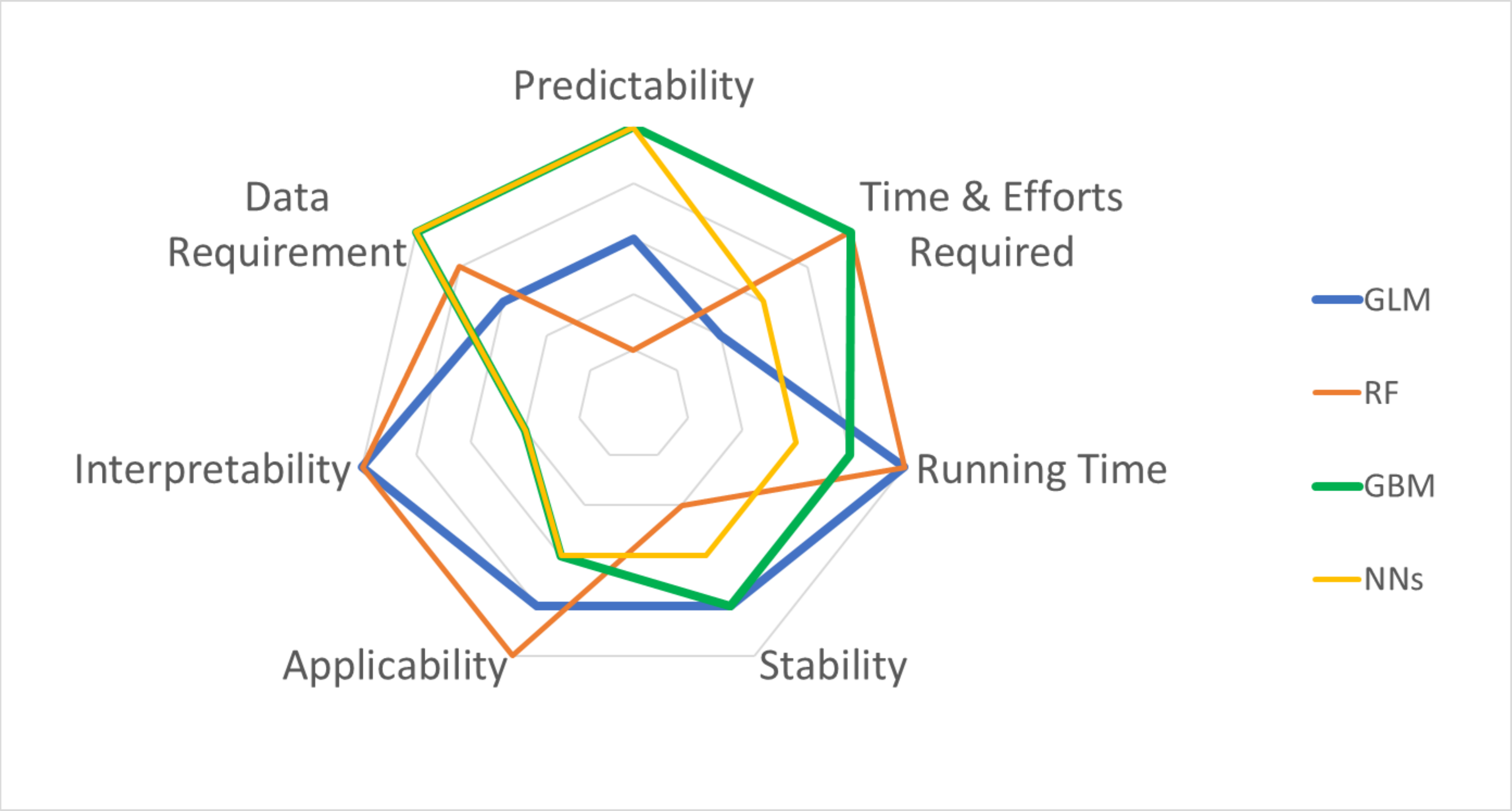
Reference:
https://www.casact.org/education/spring/2016/presentations/C-9.pdf, & Dr. Ji Yao's unpublished research.

Summary of the Study

All models are wrong, but some are useful.

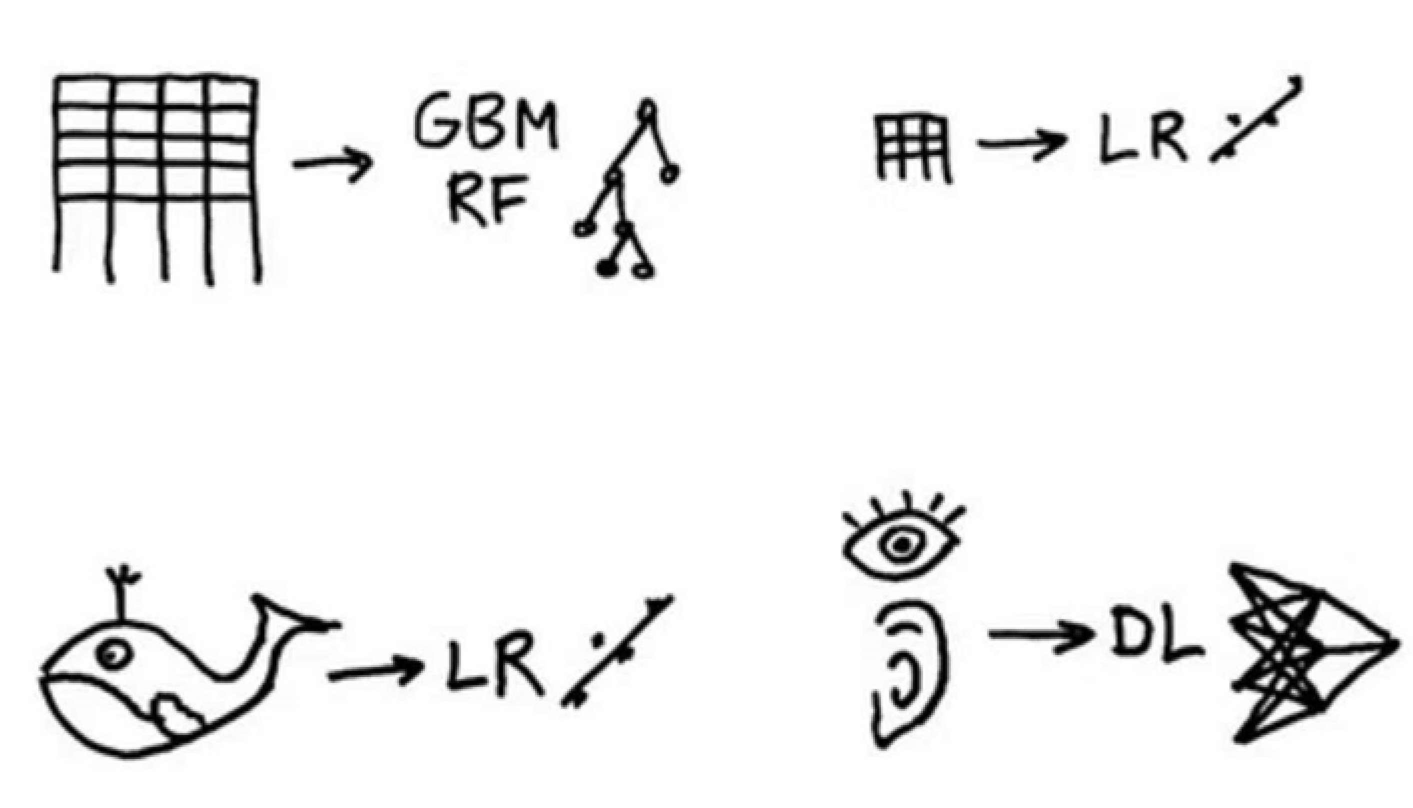
George Box

Compare the Methods for Insurance Application



Reference:
Dr. Ji Yao’s unpublished research.

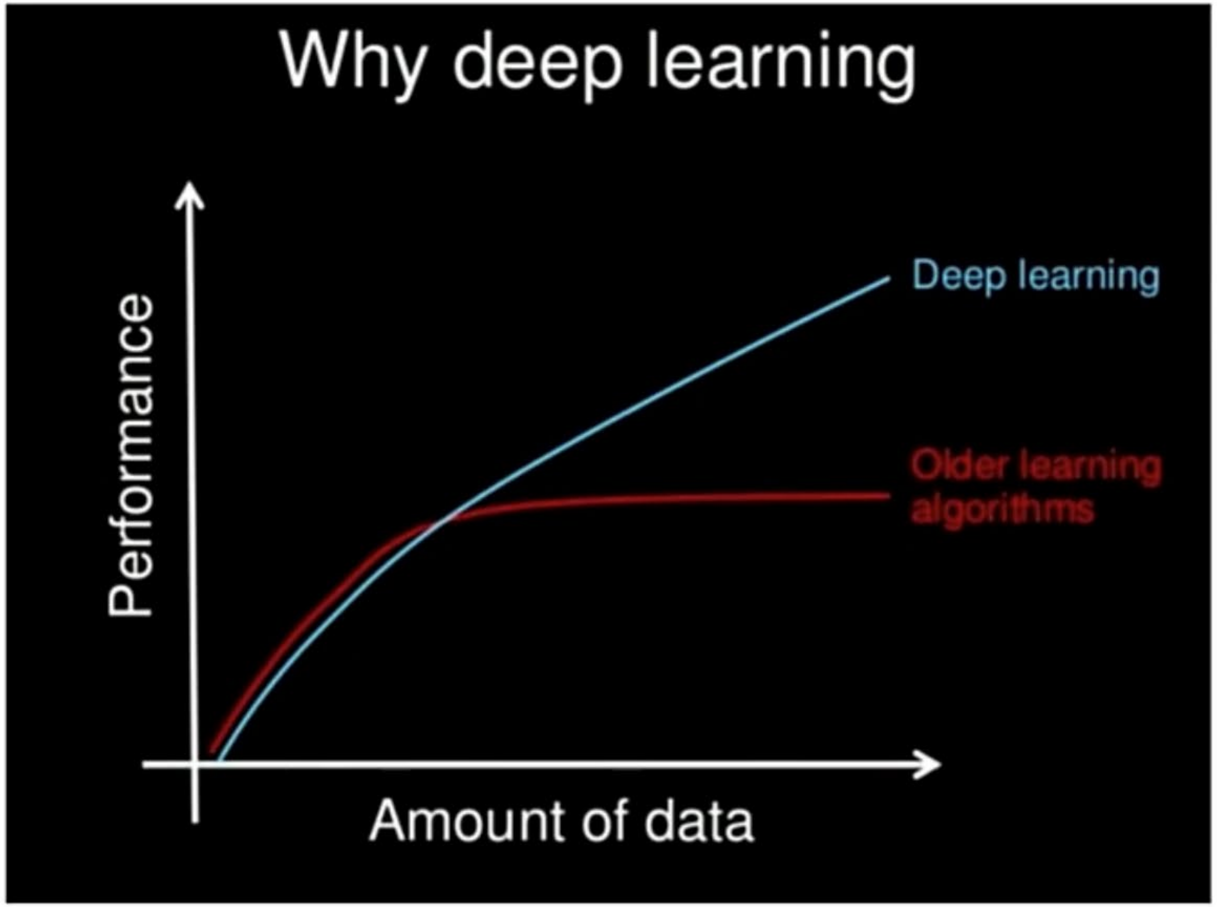
Which Algorithm is the best?



As the tasks and loss functions vary by context, the development of machine-learning methods has been relatively more problem specific.

Reference:
https://www.youtube.com/watch?v=9GCEVv94udY

Deep learning might be the next hot topic



Source: Andrew Ng

References:

[1] Christopher Cooksey, GLMs – the good, the bad, and the ugly, Midwest actuarial forum, 2009.

[2] Roel Henckaerts, etc. , Tree-based machine learning for insurance pricing, 2018.

[3] Leonardo Petrini, Non life pricing: empirical comparison of classical GLM with tree based Gradient Boosted Models, 2017.

[4] Alex Diana, etc., Machine-Learning Methods for Insurance Applications A survey, 2019.

Casualty Actuarial Society
4350 North Fairfax Drive, Suite 250
Arlington, Virginia 22203

www.casact.org