Can machine learning algorithms outperform traditional pricing methods?

Bor Harej
Bahnhofskolloquium, 6 January 2020
About me

- Actuary SAA
- 12 years at Triglav, Head of QRM
- Joined Prime Re Solutions in 2019, shareholder
- President of SAA
- Board member of ASTIN
- Champion of ASTIN working party ICDML
Agenda

- Purpose
  Comparison of quality of fit of several algorithms calibrated on synthetic policy and claim data, where “true” expected claims are known.
  
The presentation will focus on the results. The detailed explanation of individual algorithms is out of scope.

- Synthetic data
- Algorithms
- Metrics
- Results
- Overfit check
Synthetic data

Generation of policies

Motor power
Motor price
Age
Sex
Number of drivers
Distance driven

Frequency and severity distribution parameters (Poisson, Lognormal)

"True" rate (no risk loading)

Pre-defined likelihoods

Uniform likelihoods

Used to test prediction quality

One sample of portfolio claims generated (Monte Carlo)

Used for model calibration
Synthetic data

- Motor policy data (1 million)
  - Age (18-80), Sex (0,1), Motor power (1-5), Motor price (1-10), Number of drivers (1-3), owner's address (X, Y) (0-100,0-100)
  - Likelihoods defined or uniform

- Claims data
  - Poisson dist frequency, lognormal dist severity, total yearly claim used, lambda, mean and STD function of all 7 parameters
  - Function characteristics:
    - Non-linearity
    - Non-additivity
    - Non-linear dependency
    - Distance is a factor
    - Distance connected with address, some randomness added

- Simulated claims data + true policy premium rates (no risk premium)
Synthetic data: dataset overview
Synthetic data: dataset overview
Synthetic data: dataset overview
Synthetic data: rate distribution
Algorithms

- Traditionally used methods:
  - Generalized linear models (GLM)
  - Generalized additive models (GAM)

- Other machine learning algorithms:
  - Support vector machine (SVM)
  - Random forests (RF)
  - eXtreeme gradient boosting (XGB)
  - Light gradient boosting (Light GBM)
  - Neural networks (NN)
    - Regression
    - Classification
Metrics

- Charts
  - Averages by individual parameters
  - Averages by two parameters (3d)
  - Overall rate distribution

- Distance between „true“ rates and predicted rates
  - RMSE

- Market share and profit
  Assumptions:
  - Expected claim is a final premium rate
  - Cheapest option is taken
Traditionally used algorithms (GLM, GAM)
Results: GLM

- Gamma distribution, log link, all parameters used
Synthetic data: premium factors
Results: GLM

Sex is 0

Sex is 1

Simulated
GLM
Results: GAM

- 5 parameters with plain regression spline, coordinates with tensor product smooth, identity link function
Synthetic data: premium factors
Results: GAM
RESULTS

Other machine learning algorithms (RF, GB, NN)
Results: Random forests

- Max depth, number of trees, min number of leaves, max features
Synthetic data: premium factors
Results: Random forests
Results: other algorithms

XGB
- max depth, learning rate, no estimators, min child weight, gamma ...

Light GBM
- Max depth, learning rate, no estimators, boosting type, min data in leaf ...

NN
- Architecture, activation functions, optimizer, loss function, dropout rate …
- Classification classes
Results: overview

Qualitative assessment*:

<table>
<thead>
<tr>
<th></th>
<th>GLM</th>
<th>GAM</th>
<th>SVR</th>
<th>RF</th>
<th>XGB</th>
<th>Light GBM</th>
<th>NN</th>
<th>Class NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed (learning)</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Calibration complexity</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Efficiency</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

* Subjective assessment, 1 worst, 5 best

Main issue: overfitting!
Results: rate distributions
### Results: market share and profit

- **Assumption:** two players on market, one uses GLM for pricing, another classification NN

<table>
<thead>
<tr>
<th></th>
<th>GLM premium</th>
<th>Share</th>
<th>True premium</th>
<th>Claims</th>
<th>Profit</th>
<th>Relative profit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total premium</strong></td>
<td>241.323.027</td>
<td>100%</td>
<td>238.995.840</td>
<td>244.783.691</td>
<td>-3.460.664</td>
<td>-1,4%</td>
</tr>
<tr>
<td><strong>Premium below market price</strong></td>
<td><strong>65.323.200</strong></td>
<td>27%</td>
<td>140.170.056</td>
<td>145.159.954</td>
<td>-79.836.754</td>
<td>-122,2%</td>
</tr>
<tr>
<td><strong>Premium above market price</strong></td>
<td>175.999.827</td>
<td>73%</td>
<td>98.825.783</td>
<td>99.623.738</td>
<td>76.376.089</td>
<td>43,4%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Class NN premium</th>
<th>Share</th>
<th>True premium</th>
<th>Claims</th>
<th>Profit</th>
<th>Relative profit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total premium Class NN</strong></td>
<td>229.582.146</td>
<td>100%</td>
<td>238.995.840</td>
<td>244.783.691</td>
<td>-15.201.545</td>
<td>-6,6%</td>
</tr>
<tr>
<td><strong>Premium below market</strong></td>
<td><strong>95.056.387</strong></td>
<td>41%</td>
<td>98.825.783</td>
<td>99.623.738</td>
<td>-4.567.351</td>
<td>-4,8%</td>
</tr>
<tr>
<td><strong>Premium above market</strong></td>
<td>134.525.759</td>
<td>59%</td>
<td>140.170.056</td>
<td>145.159.954</td>
<td>-10.634.195</td>
<td>-7,9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Total market premium</th>
<th>160.379.587</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Winner's curse ratio</strong></td>
<td>67,1%</td>
<td></td>
</tr>
</tbody>
</table>
# Results: Market Share and Profit

<table>
<thead>
<tr>
<th></th>
<th>Total claims</th>
<th>Total premium</th>
<th>Total premium GLM</th>
<th>Total premium RF</th>
<th>Total premium Light GBM</th>
<th>Total premium XGB</th>
<th>Total premium GAM</th>
<th>Total premium NN</th>
<th>Total premium Class NN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total claims</strong></td>
<td>244,783,691</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total premium</strong></td>
<td>238,995,840</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total premium GLM</strong></td>
<td>241,323,027</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total premium RF</strong></td>
<td>243,907,662</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total premium Light GBM</strong></td>
<td>243,119,068</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total premium XGB</strong></td>
<td>244,175,522</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total premium GAM</strong></td>
<td>243,540,112</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total premium NN</strong></td>
<td>237,157,861</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total premium Class NN</strong></td>
<td>229,582,146</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Relative Profit Ratio

<table>
<thead>
<tr>
<th>party/market</th>
<th>GLM</th>
<th>GAM</th>
<th>RF</th>
<th>XGB</th>
<th>Light GBM</th>
<th>NN</th>
<th>Class NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLM</td>
<td>-86.4%</td>
<td>-101.3%</td>
<td>-105.7%</td>
<td>-108.1%</td>
<td>-94.1%</td>
<td>-122.2%</td>
<td></td>
</tr>
<tr>
<td>GAM</td>
<td>-6.9%</td>
<td>-32.7%</td>
<td>-31.0%</td>
<td>-34.9%</td>
<td>-21.1%</td>
<td>-43.8%</td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td>1.5%</td>
<td>-9.0%</td>
<td>-12.3%</td>
<td>-15.7%</td>
<td>-10.3%</td>
<td>-21.3%</td>
<td></td>
</tr>
<tr>
<td>XGB</td>
<td>1.5%</td>
<td>-1.6%</td>
<td>-9.8%</td>
<td>-10.3%</td>
<td>-1.8%</td>
<td>-15.8%</td>
<td></td>
</tr>
<tr>
<td>Light GBM</td>
<td>2.7%</td>
<td>-3.7%</td>
<td>-3.6%</td>
<td>-4.5%</td>
<td>-3.1%</td>
<td>-11.8%</td>
<td></td>
</tr>
<tr>
<td>NN</td>
<td>-15.9%</td>
<td>-13.8%</td>
<td>-26.2%</td>
<td>-25.7%</td>
<td>-27.8%</td>
<td>-33.2%</td>
<td></td>
</tr>
<tr>
<td>Class NN</td>
<td>-4.8%</td>
<td>-8.9%</td>
<td>-13.2%</td>
<td>-12.6%</td>
<td>-14.5%</td>
<td>-6.8%</td>
<td></td>
</tr>
</tbody>
</table>
### Overfit check

- Re-run policy and claim simulations and use calibrated models for prediction

<table>
<thead>
<tr>
<th>RMSE</th>
<th>(in 1.000 units)</th>
<th>Calibration</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLM total predictions</td>
<td>296</td>
<td>296</td>
<td></td>
</tr>
<tr>
<td>GAM total predictions</td>
<td>211</td>
<td>212</td>
<td></td>
</tr>
<tr>
<td>Random forest predictions</td>
<td>202</td>
<td>203</td>
<td></td>
</tr>
<tr>
<td>XGBoost predictions</td>
<td>195</td>
<td>196</td>
<td></td>
</tr>
<tr>
<td>Light GBM predictions</td>
<td>194</td>
<td>195</td>
<td></td>
</tr>
<tr>
<td>Neural networks predictions</td>
<td>205</td>
<td>206</td>
<td></td>
</tr>
<tr>
<td>Classification neural networks</td>
<td>193</td>
<td>194</td>
<td></td>
</tr>
</tbody>
</table>
Summary

- Traditionally pricing methods can be outperformed
  - Example shown on synthetic data

- Different algorithms, might be tricky to calibrate and not to overfit

- Best fits: Light GBM, Classification NN

- Strong effect on profitability and market share

- Results seem to be stable
Bor Harej
bor.harej@prs-zug.com

QUESTIONS?