

AIcrowd

Motor insurance market simulation

Imperial College London · Singapore Actuarial Society · Actuaries Institute Australia · CAS · Institute and Faculty of Actuaries · IAA AAI ASTIN Non-Life Insurance · UQÀM

Prizes $12,000 (USD)

Michaël Bordeleau

March 21st, 2021

HI!

- Montréal, Québec, Canada
- « Typical » Actuary
- Home & Auto Insurance
- working with GLMs



Towers Watson
Emblem



sas

HI (PT.2)!

- Working within a Research & Development team
- Young Data Scientists
- All skillful in Python and R
- Have been watching the parade from the sidelines…



- This competition was my motivation at **learning R** and initiate a journey into **machine learning**











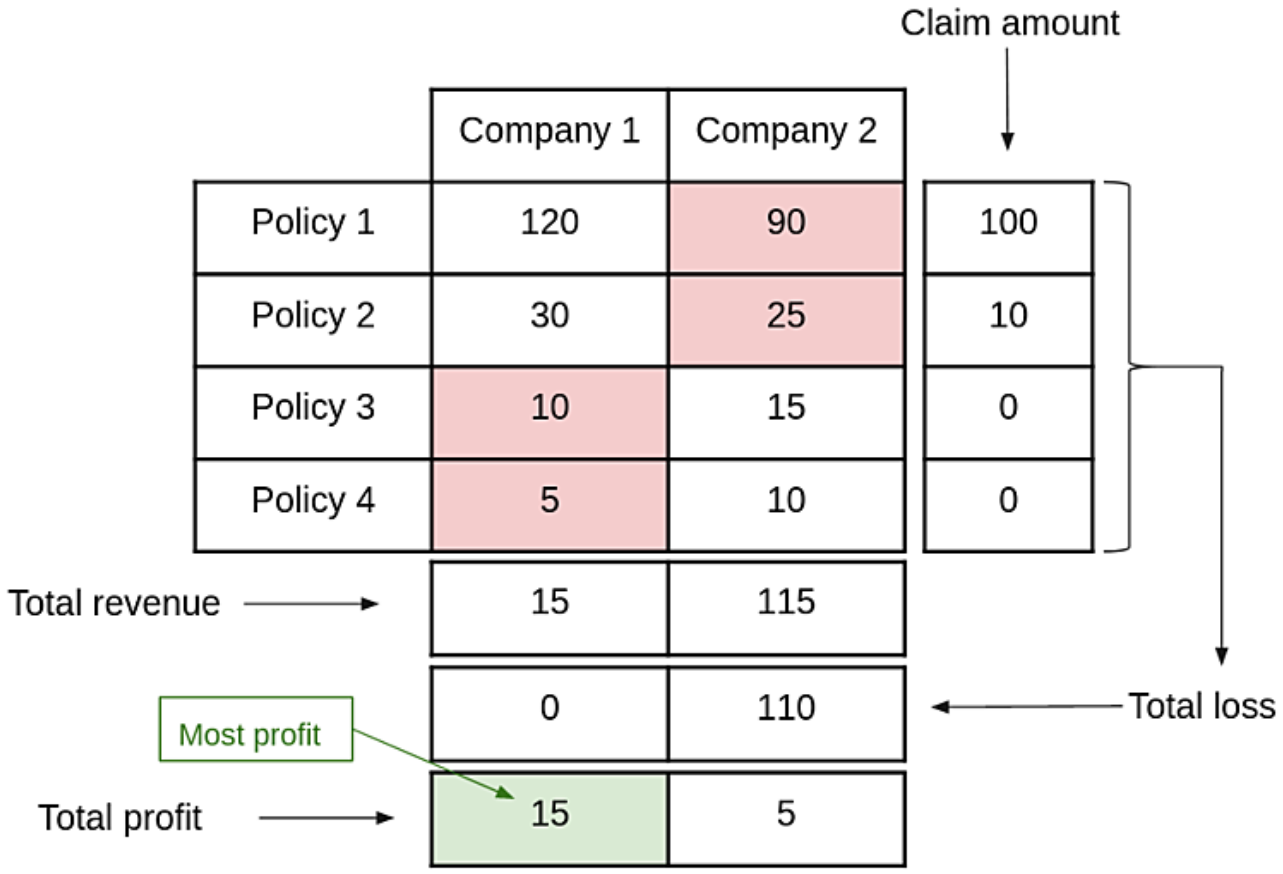
COMPETITION

GOAL

You are an insurance company, building a pricing model, and compete against other insurers for profit.

Clients pick lowest price quote.

Insurer with most profit wins.



DATA

	POLICY	DRIVER(S)	VEHICLE	GEOGRAPHY
id_policy	coverage	gender	make model (hash)	population
year (1-4)	age	age	age	town surface area
claim amount	payment frequency	years licensed	usage type	
	ppv vs commercial	no claims discount	fuel	
	telematics (y/n)		speed	
			weight	
			value (as new)	

- From an European insurer
- 57,054 policies x 4 years = 228,216 rows
- Note that there's no variable tracking past claim counts and amounts

FEATURE ENGINEERING

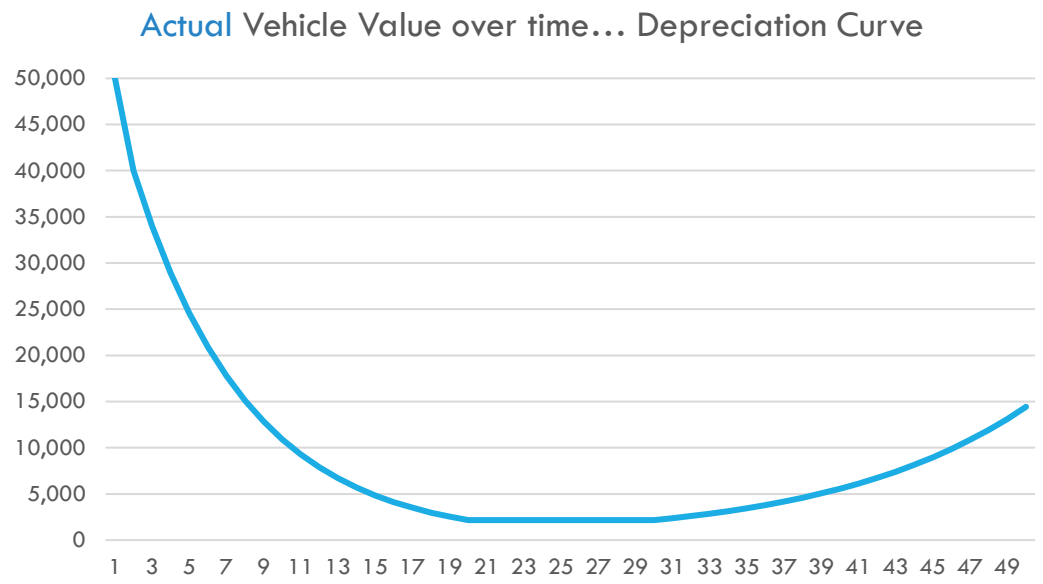
Go crazy



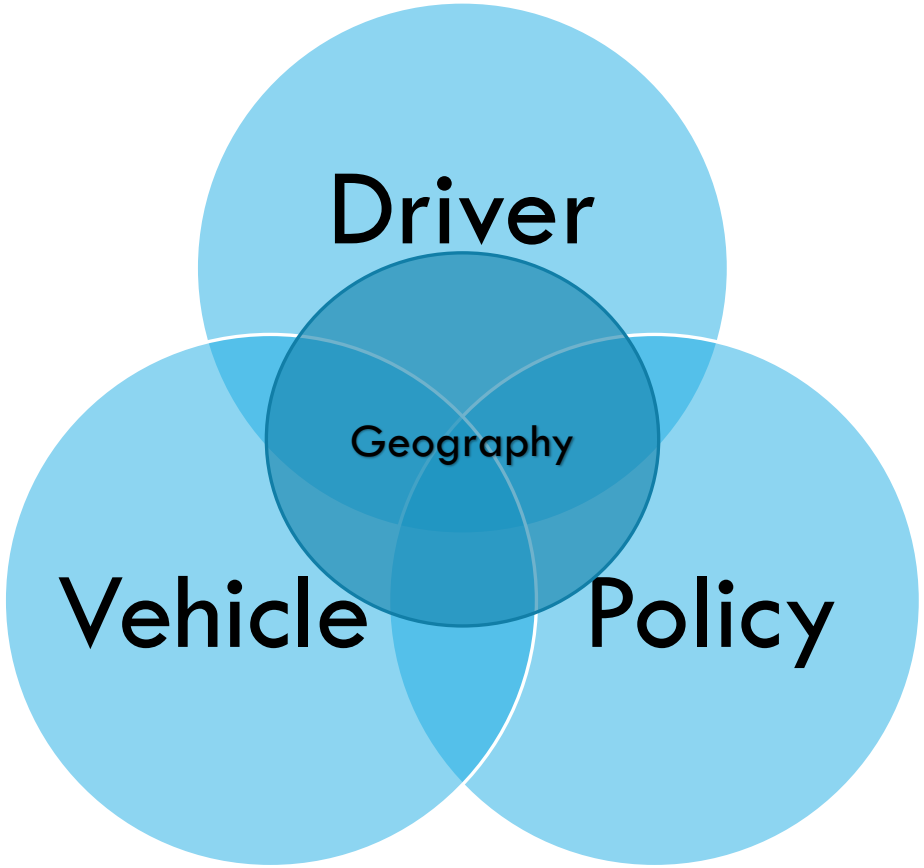
TRANSFORMATIONS

- Population Density = Population / Surface Area

- Actual Vehicle Value :: exponential depreciation
  - with reversal for very old vehicles, “collectibles”

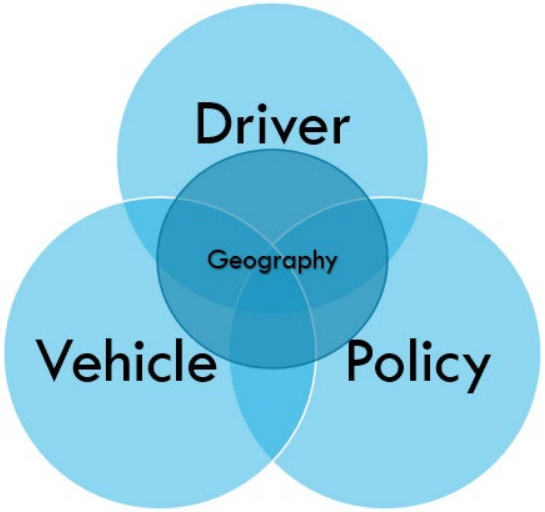


INTERACTIONS



thorough **thinking** about the variables and how it could influence the risk

INTERACTIONS :: ASSUMPTIONS



- Young/Inexperienced driver with a faster vehicle can be more risky
- Young/Inexperienced driver with a heavier vehicle can be more risky
- Young/Inexperienced driver with an expensive vehicle can be more risky
- Young/Inexperienced driver in a densified city can be more risky

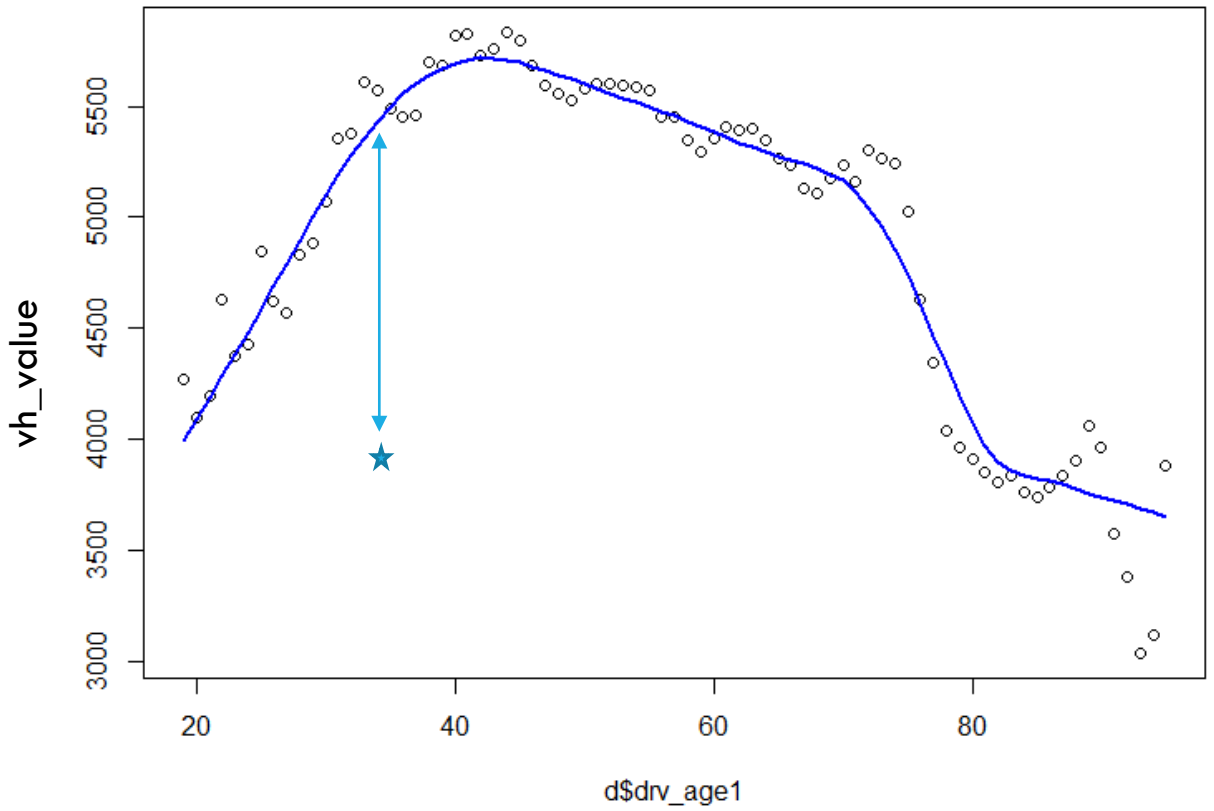
- or brute force :: create all possible ratio combinations
  - Variables x1, x2, x3, … xn.
  - x1/x2 , x1/x3 , x1/xn , … x2/x3 , x3/x4 , …
  - x1- x2 , x1- x3 , x1- xn , … x2- x3 , x2- x4 , …

YOU VS THE WORLD

★ Are **you** driving a vehicle that is much more **expensive**/ or less than your friend?

Measure the distance between:
. your vehicle value
. the smoothed average vehicle value (blue line), at your age

Reflects client behavior, frugality?

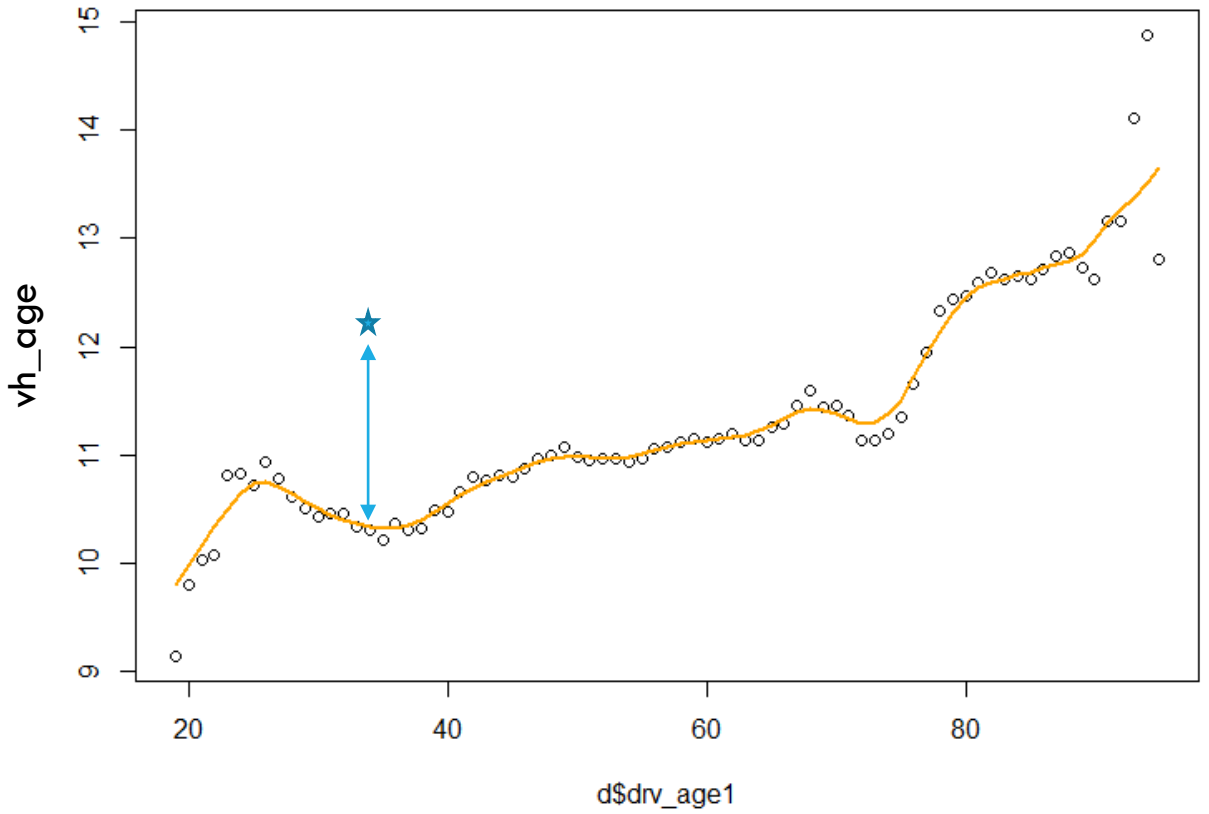


YOU VS THE WORLD

- Are **you** driving a much **newer** or **older vehicle** than your friend?

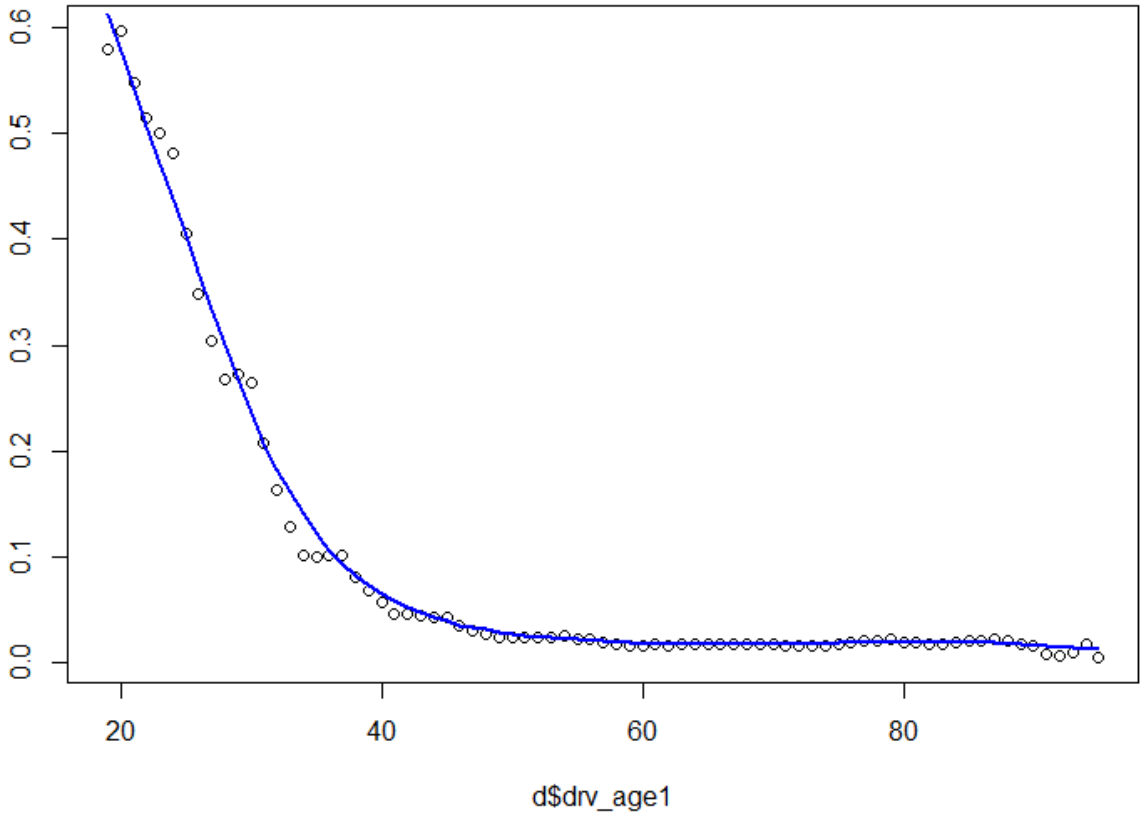
Made selections on smoothing coefficient

And so on…



YOU VS THE WORLD

Do you have a much better or worse **No Claims Discount** than your friend?

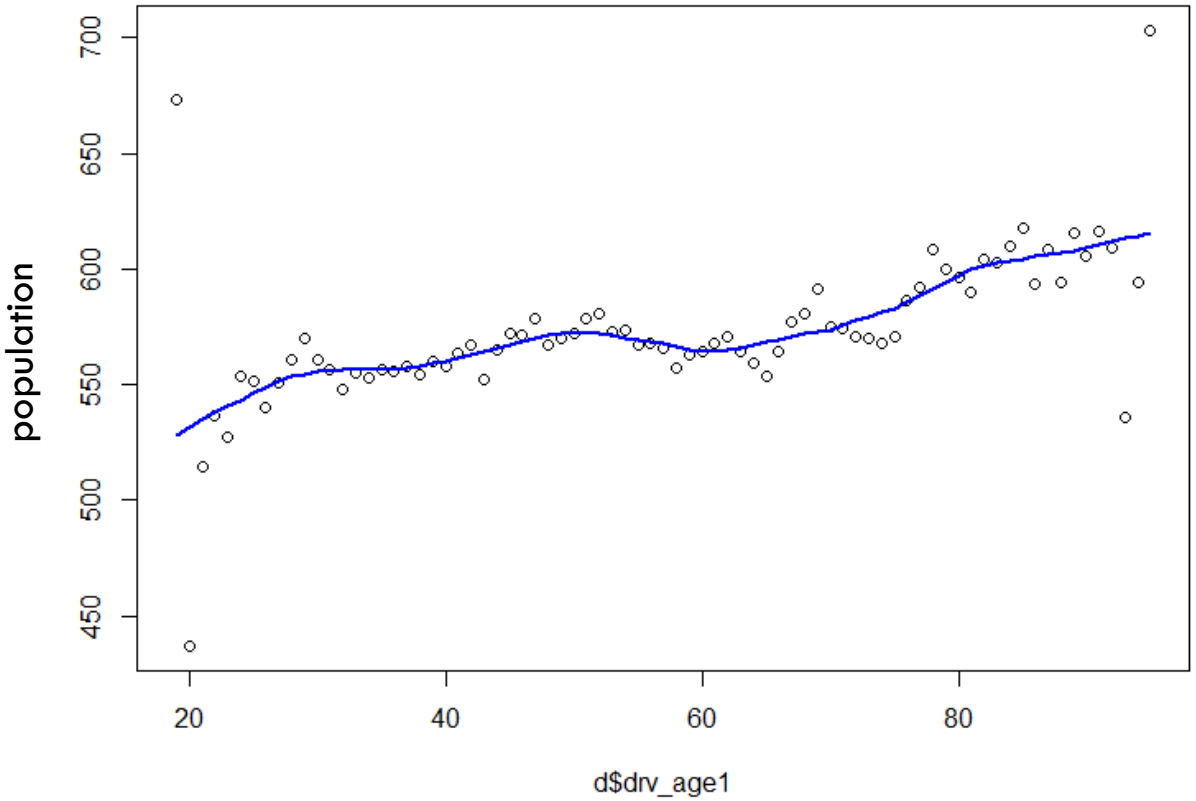


YOU VS THE WORLD

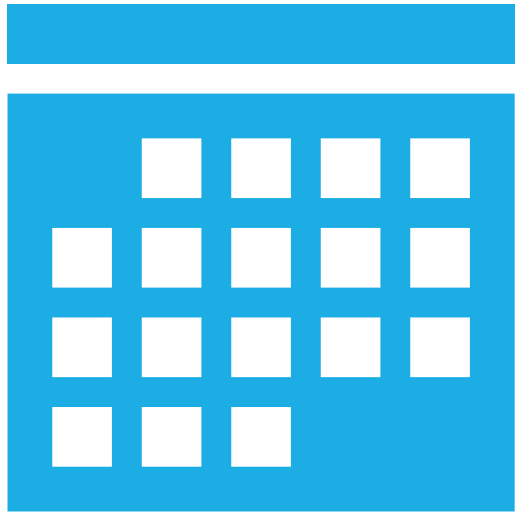
Are you living in a much more **populous city** than your friend?

… and so on.

Repeated same exercise by Driver Experience (Number of Years Licensed)



I KNOW WHAT YOU DID LAST SUMMER



- Created features based on **your own history** with the insurer

  - Have you switched vehicle?
  - Have you moved to a new city?
  - Have you modified your insurance coverage?
  - Have you changed your payment frequency?
  - How has your No Claims Discount evolved?

  - Did you claim and how much money have I been giving you?!!?

GET SOME EDGE?



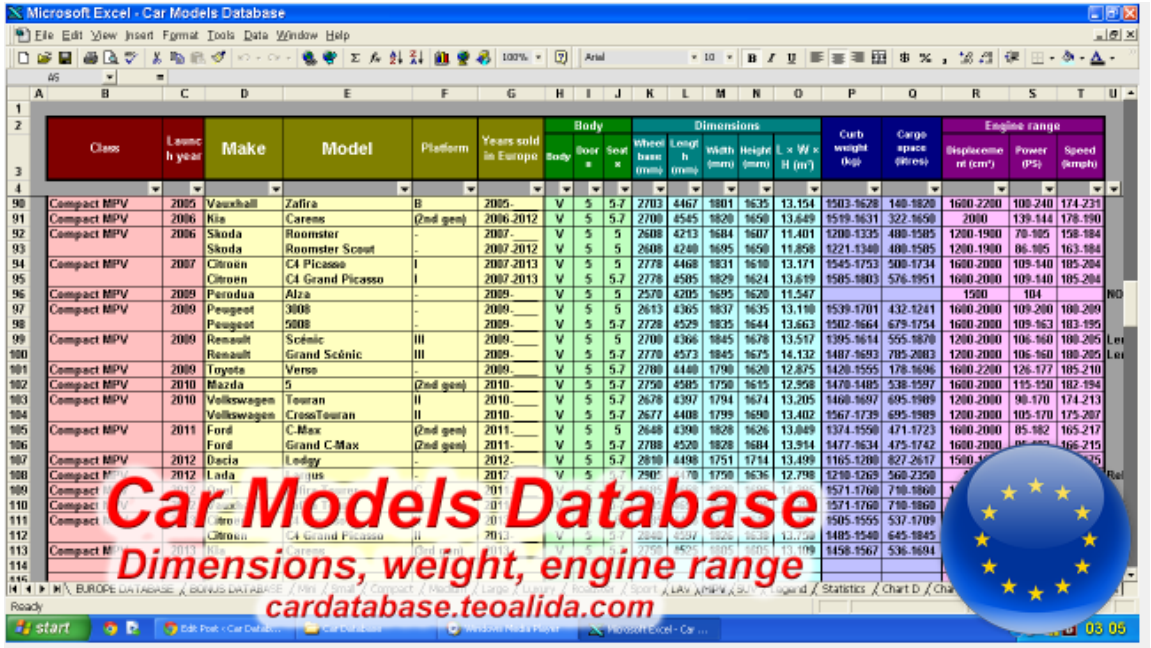
External Database

Merge using Speed & Weight

Cross-check reasonableness by comparing popularity of vehicle

Get engine specs, …

Profit?



$110 USD

CATEGORICAL VARIABLES :: ONE-HOT ENCODING



- Policy Coverage has 4 levels
  - Max
  - Med2
  - Med1
  - Min

- Replace with 4 binary variables:
  - pol_coverage.max 0 or 1
  - pol_coverage.med2 0 or 1
  - pol_coverage.med1 0 or 1
  - pol_coverage.min 0 or 1

- For models sensitive to collinearity (ie. GLMs), preferable to leave one out
  - Avoids the “Dummy Variable Trap”

CATEGORICAL VARIABLES :: TARGET ENCODING



- Use insights from the response variable (claim amount) to transform into a numerical value
- Policy Coverage has 4 levels
  - Max
  - Med2
  - Med1
  - Min
- In terms of frequency and severity of claims,
  - Max looked higher risk than Med2, which was riskier than Med1 which was riskier than Min.
- Replace with a numerical value* correlated with the risk.
  - Max ➔ 4
  - Med2 ➔ 3
  - Med1 ➔ 2
  - Min ➔ 1

CATEGORICAL VARIABLES :: TARGET ENCODING



You can target encode using various type of values:

- Mean
- Median
- Order

- However, beware of …

LEAKAGE

Garbage in, Garbage out



LEAKAGE



- As I progressed in the competition, I saw my public RMSE score deteriorate
- If you do too much target encoding, you WILL experience leakage
- Model performs great in cross-validation but will be useless on new data

LEAKAGE

- For Vehicle Make Model, I did some mean target encoding
  - with heavy smoothing
  - still ended up having leakage

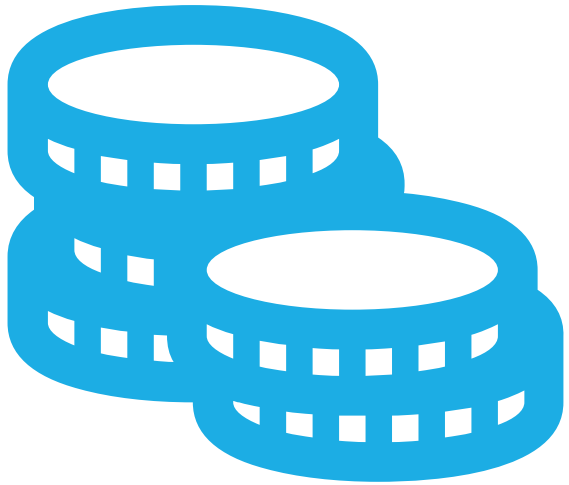
- Ultimately, the only feature I kept was a Vehicle "Popularity" index
  - Quantify the "Popularity" of a Vehicle in the population. Heavy rounding.
  - Index ranging from 1 to 78

- Rational:
  - A Popular Vehicle could be cheaper to repair due to broader parts availability, and vice-versa



CLAIM AMOUNT, YEAR AND INFLATION

- Originally had capped losses at $45,000
- No effort was put into analyzing large losses



- Selected inflation rate per year
- Inflated all losses up to $T = 4$
- Dropped the year variable

- I didn't want year to be used as a feature in any of my models

MODELING

You get a model, you get a model, everyone gets a model

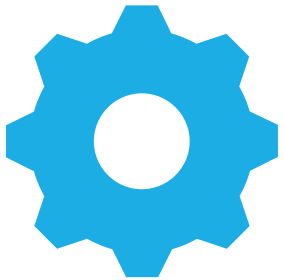


MODELS

18 different models to tune

12 different prediction models

- xgboost
- catboost
- lightgbm
- random forest
- glm
- glmnet

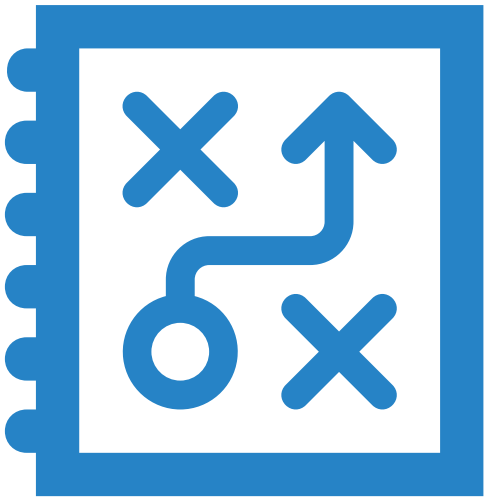


- frequency x severity
- loss cost (tweedie)

MODELS

- For GLMs, dimensionality reduction using Principal Components Analysis

- Hyperparameters Tuning using Grid Search and Parallel Processing





STACKING

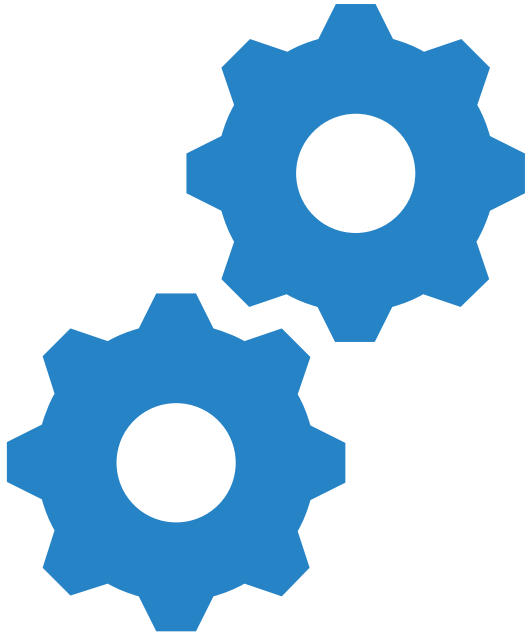
. To improve claims prediction

. 4 meta-learning models

. Trained on top of all the models developed

- . xgboost
- . lightgbm
- . random forest
- . neural network

. Final is an average of the 4



18 base models
+ 4 meta models

22 models tuned



MARKET SIMULATION

Another way to heat up your house in winter!



MARKET SIMULATION

- Using 11 models, trained 8 different times (new seeds)
- 8 different 70 train / 30 hold out.
- There's no model winning systematically

Profit Rank

		S_17	S_42	S_666	S_1313	S_4862	S_90210	S_9999	S_9990	Avg		Std
A	rf_fxs	4	1	1	3	2	6	1	1	2.4	A	1.7
B	xgb_fxs	6	2	6	1	4	2	4	4	3.6	B	1.7
C	lgb_fxs	3	7	3	2	3	4	6	5	4.1	C	1.6
D	glm_fxs	2	9	2	7	5	8	5	3	5.1	D	2.5
E	cat_fxs	5	3	9	5	7	7	3	2	5.1	E	2.3
F	glm_lcost	1	6	4	6	10	3	7	6	5.4	F	2.5
G	lgb_lcost	9	8	8	10	6	1	2	8	6.5	G	3.1
I	rf_lcost	8	4	7	9	1	9	8	9	6.9	I	2.7
J	xgb_lcost	7	5	11	4	9	5	10	7	7.3	J	2.4
K	cat_lcost	10	10	5	8	8	10	9	10	8.8	K	1.6
L	glmnet_lcost	11	11	10	11	11	11	11	11	10.9	L	0.3
												2.0

MARKET SIMULATION

Understanding loss functions
and its correlation with profit

RMSE test rank

		S_17	S_42	S_666	S_1313	S_4862	S_90210	S_9999	S_9990	Avg		Std
A	rf_fxs	9	4	8	9	8	9	9	3	7.4	A	2.3
B	xgb_fxs	4	2	1	1	1	1	1	2	1.6	B	1.0
C	lgb_fxs	3	3	3	3	2	2	4	4	3.0	C	0.7
D	glm_fxs	1	11	11	4	10	3	7	11	7.3	D	3.8
E	cat_fxs	6	1	2	2	3	7	2	1	3.0	E	2.1
F	glm_lcost	2	10	10	5	11	4	11	10	7.9	F	3.4
G	lgb_lcost	8	7	4	11	7	6	5	8	7.0	G	2.0
I	rf_lcost	11	6	6	6	4	10	3	6	6.5	I	2.5
J	xgb_lcost	5	5	5	7	5	5	10	5	5.9	J	1.7
K	cat_lcost	10	8	9	10	6	11	8	9	8.9	K	1.5
L	glmnet_lcost	7	9	7	8	9	8	6	7	7.6	L	1.0
												2.0

GINI test rank

		S_17	S_42	S_666	S_1313	S_4862	S_90210	S_9999	S_9990	Avg		Std
A	rf_fxs	10	5	7	7	1	9	7	4	6.3	A	2.7
B	xgb_fxs	5	1	4	1	4	1	3	3	2.8	B	1.5
C	lgb_fxs	3	2	6	3	6	2	5	1	3.5	C	1.8
D	glm_fxs	1	10	2	5	10	8	10	8	6.8	D	3.4
E	cat_fxs	4	4	1	2	7	7	2	2	3.6	E	2.2
F	glm_lcost	2	9	3	4	11	5	11	9	6.8	F	3.4
G	lgb_lcost	8	7	5	6	3	3	1	7	5.0	G	2.3
I	rf_lcost	11	8	10	11	8	11	6	11	9.5	I	1.8
J	xgb_lcost	6	6	9	8	5	4	9	5	6.5	J	1.8
K	cat_lcost	9	3	11	9	2	6	4	10	6.8	K	3.2
L	glmnet_lcost	7	11	8	10	9	10	8	6	8.6	L	1.6
												2.3

MARKET SIMULATION

Understanding that models have natural tendency to win different segments

Average Premium on policies "won"

		S_17	S_42	S_666	S_1313	S_4862	S_90210	S_9999	S_9990	Avg	Std
A	rf_fxs	173	174	175	175	177	173	172	177	175	1.8
B	lgb_fxs	125	129	127	129	119	125	134	120	126	4.5
C	xgb_fxs	114	119	111	120	114	109	114	119	116	3.7
D	glm_fxs	119	119	119	117	119	117	118	119	119	0.8
E	glm_lcost	129	128	128	126	127	128	129	130	127	1.2
F	cat_fxs	150	152	153	146	150	151	154	147	150	2.5
G	rf_lcost	118	117	115	117	123	120	122	123	118	2.7
I	xgb_lcost	107	126	108	111	110	116	111	113	112	5.7
J	lgb_lcost	98	86	88	93	93	91	83	91	92	4.4
K	cat_lcost	93	100	100	99	98	101	97	96	98	2.6
L	glmnet_lcost	75	73	73	73	74	74	76	72	74	1.3
											2.8



PROFIT LOADING

Bringing home the bacon!

TIME WASTED…

- Spent 10 weeks working on:
  - Features
  - Models
  - Simulation
- Spent 2 minutes on profit loading
- Had no idea how to adequately load my prices
- Didn't find much value in the Weekly Feedback beside Market Share %
- Had to **improvise**

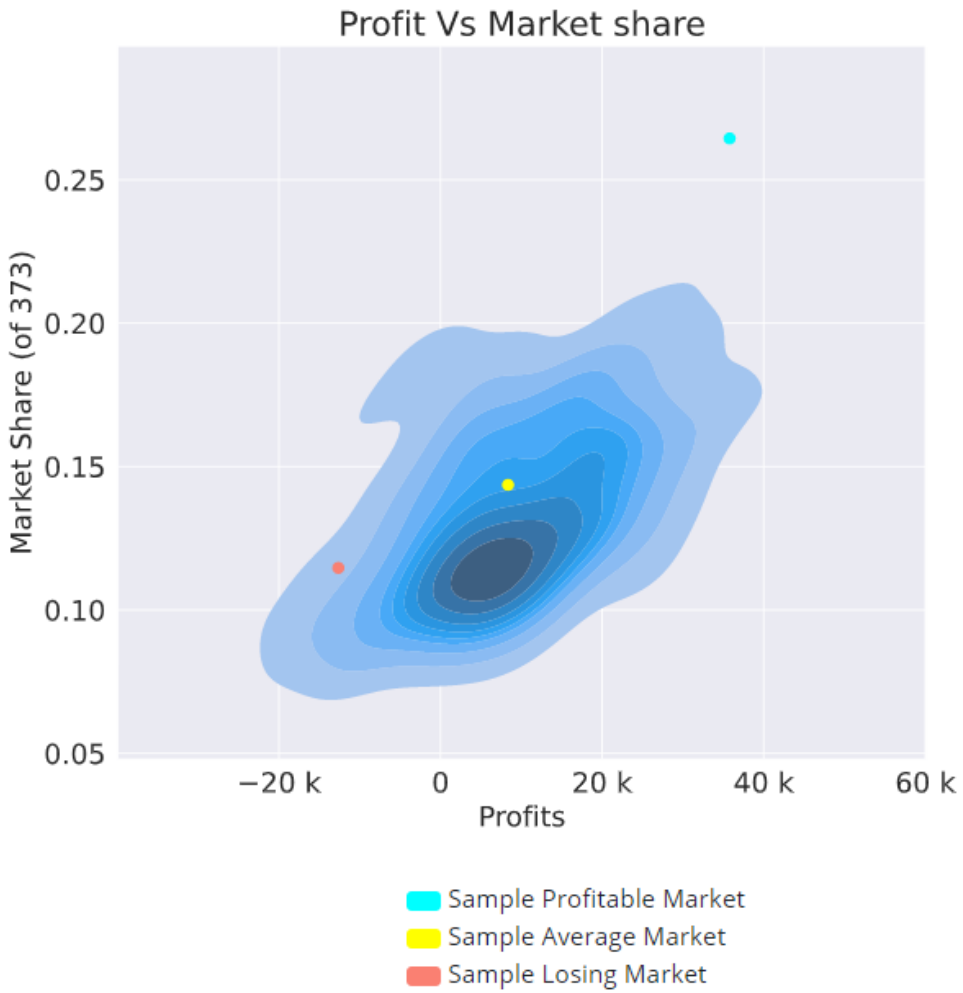


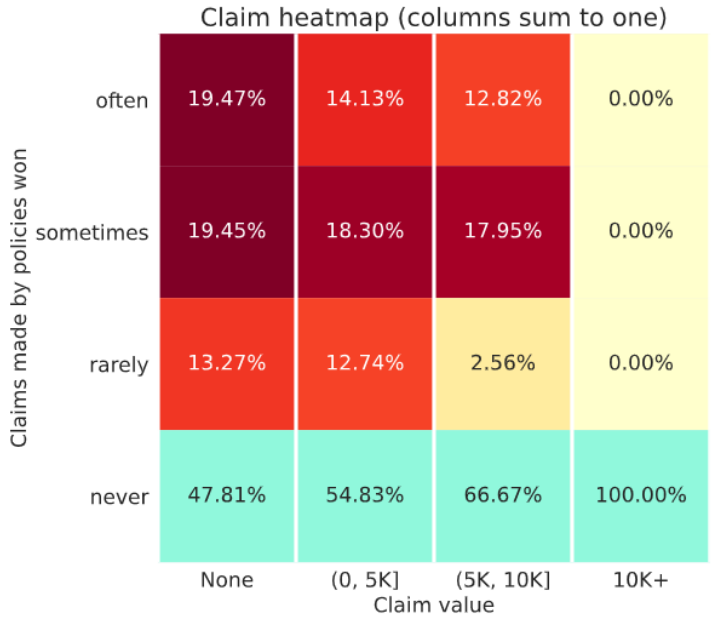
LEADERBOARD PROGRESS

Week	Teams	Position	Profit	Market Share
1	44	--	--	--
2	81	67	-271,294	44 %
3	107	107	Overcharged	--
4	142	90	-118,561	20 %
5	168	14	-323	0.4 %
6	191	8	6,420	19 %
7	207	61	-53,298	18 %
8	220	43	-11,015	17 %
9	247	99	-14,548	11 %
10	259	8	9,296	13 %

Each week is independent

FEEDBACK (WEEK 10)





Average Market Share	0.134
Average Revenue	193569.0
Average Losses	184273.0
Average Loss Ratio	0.957
Average Profit per Policy	3.0
Number of Markets	373
Average Profit	9296.4

MINIMUM LOADING TO ACHIEVE PROFITABILITY?

- Recycling market simulation, building markets of 10 “insurers”
- Recursively identified the minimal loading required
  - so that *all 10 insurers* end up profitable
- Highly volatile results… between 25% and 40% loading required
- Looks high… I should aim a little lower
  - I need to expect, *and desire*, that some insurers will lose money…

PROFIT LOADING

Two loadings:

.Variable Loading.
proportional to the expected claim

$Vc = 1.149$

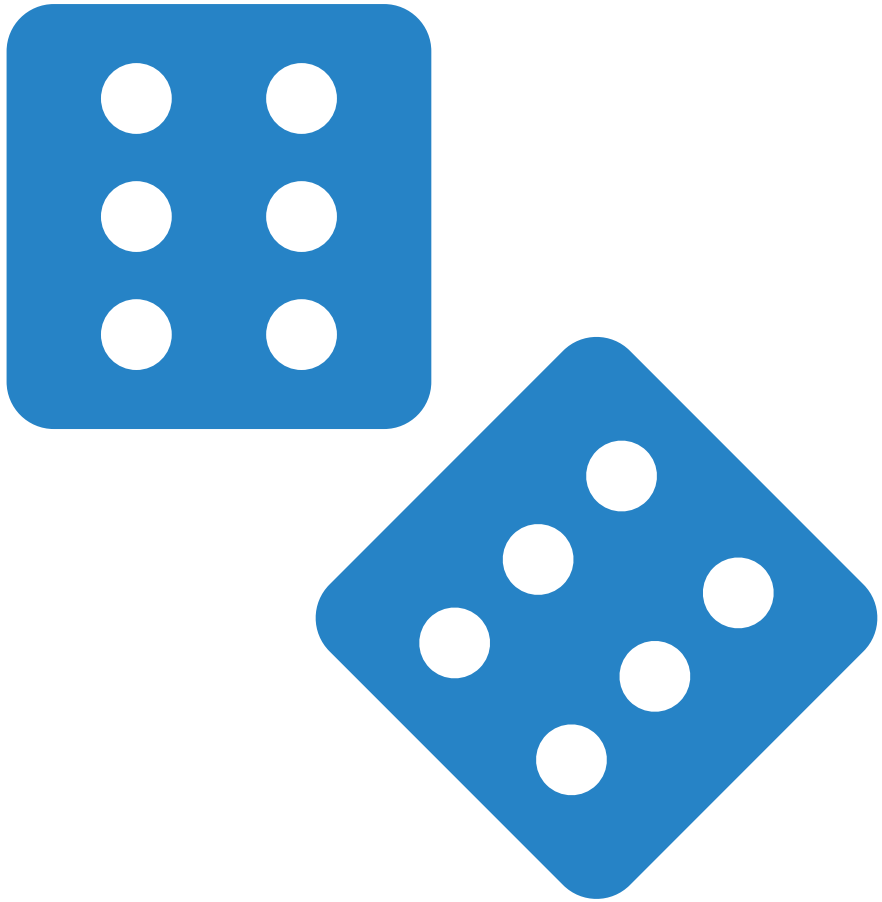
.Variable Loading.
partially proportional to the *odds of having* a claim

$Vf = 1 + E[\text{frequency}] \times 63\%$

Premium = E[Claims] x Vc x Vf

Loading distribution:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.154	1.201	1.230	1.228	1.253	1.594



END GAME

- Quote 100,000 policies at year 5
  - 43 % are new clients
  - 57 % are clients from the training data (year 1-4)
- Submitted a zip file hosting 2 models:
  - Model A : generalized model fitting to all business*
  - Model B : data-augmented model fitting to known clients
    - saved past claim counts & amounts from the training data
    - fed it to the quoting process
- Cherry picking 🍒



RESULTS



252 teams

between 300 and 400 participants

43e Simon Coulombe

37e Sylvain Demars

21e Maxime Comeau

4e Michaël Bordeleau

2e Guillaume Lepage


THANK YOU
Questions?
linkedin.com/in/mbordeleau