Synthetic Dataset Generation of Driver Telematics

Banghee So, Jean-Philippe Boucher, and Emiliano A. Valdez
University of Connecticut and Université du Québec à Montréal (UQAM)

Insurance Data Science Conference
City, University of London
Online 16-18 June 2021 GMT+1

Background on telematics

- Telematics is about use of telecommunication devices and technology to transmit and store information.
- Wide applications in the automobile industry.
- An intelligent device may be installed in the car to: (a) monitor and transmit driving information; (b) remotely control driverless cars; (c) monitor fleets for a systematic and efficient manner (e.g., Uber)
- It enables insurers to collect driving metrics to enhance driver risk profile.
- Economic benefits: studies show drivers save about 5-15% annually with UBI.
    - Social benefits e.g., reduces congestion, car emissions

UCONN

Synthetic dataset generation

- Motivation: why?
- Synthetic data: what?
- The procedure: how?
- The output: how well?
- Where to acquire data?

Overview of existing work

Data source	Reference	Sample	Period	Analytical techniques	Research synthesis
Belgium	Verbelen et al. (2018)	10,406 drivers (33,259 obs.)	2010-2014	Poisson GAM, Negative binomial GAM	Shows that the presence of telematics variables are better important predictors of driving habits
China	Gao et al. (2019)	1,478 drivers	2014.01-2017.06	Poisson GAM	Shows the relevance of telematics covariates extracted from speed-acceleration heatmaps in a claim frequency model
Europe	Baecke & Bocca (2017)	6,984 drivers ($<$ age 30)	2011-2015	Logistic regression, Random forests, Neural networks	Illustrates the importance of telematics variables for pricing UBI products and shows that as few as three months of data may already be enough to obtain efficient risk estimates
Greece	Guillen et al. (2020)	157 drivers (1,225 obs.)	2016-2017	Negative binomial reg.	Demonstrates how the information drawn from telematics can help predict near-miss events
Japan	Osafune et al. (2017)	809 drivers	2013.12-2015.02	Support Vector Machines	Investigates accident risk indices that statistically separate safe and risky drivers
Spain	Ayuso et al. (2014)	15,940 drivers ($<$ age 30)	2009-2011	Weibull regression	Compares driving behaviors of novice and experienced young drivers with PAYD policies
	Ayuso et al. (2016)	8,198 drivers ($<$ age 30)	2009-2011	Weibull regression	Determines the use of gender becomes irrelevant in the presence of sufficient telematics information
	Boucher et al. (2017)	71,489 obs.	2011	Poisson GAM	Offers the benefits of using generalized additive models (GAM) to gain additional insights as to how premiums can be more dynamically assessed with telematics information
	Guillen et al. (2019)	25,014 drivers ($<$ age 40)	2011	Zero-inflated Poisson	Investigates how telematics information helps explain part of the occurrence of zero accidents not typically accounted by traditional risk factors
	Ayuso et al. (2019)	25,014 drivers ($<$ age 40)	2011	Poisson regression	Incorporates information drawn from telematics metrics into classical frequency model for tariff determination
	Perez-Marin et al. (2019)	9,614 drivers ($<$ age 35)	2010	Quantile regression	Demonstrates that the use of quantile regression allows for better identification of factors associated with risky drivers
	Pesantez-Narvaez et al. (2019)	2,767 drivers ($<$ age 30)	2011	XGBoost	Examines and compares the performance of XGBoost algorithm against the traditional logistic regression

Motivation: why?

- Despite growth in actuarial/insurance literature, there is limited accessibility to driver telematics dataset.
  - Concern of confidentiality and privacy
- In actuarial, there is need for datasets to conduct risk assessments with high accuracy and efficiency.
  - Involves construction, calibration, testing big and diverse data with meaningful information.
  - Other disciplines have similar need for flexible/rich datasets to conduct tests of various algorithms in machine learning.
- We follow a trend in interest and demand for a repository of data to perform insurance analytics:
  - Synthetic datafile on a large portfolio of VA products: Gan and Valdez (2017, 2018)
  - Stochastically simulated insurance data intended for granular data reserving: Gabrielli and Wuthrich (2018)
  - Similar trend in other disciplines, e.g., medicine: Synthetic patient data, Goncalves, et al. (2020)

UCONN

Synthetic dataset: what?

- Features of the synthetic dataset:
  - 100,000 policies are synthetically generated
  - The file was an imitation from a real empirical dataset on driver telematics (static, not dynamic)
- Empirical data from Canadian-owned company offering insurance and investment products:
  - UBI auto program was first launched in year 2013
- Observation period: 2013–2016

Description of variables in the synthetic dataset

Type	Variable	Description
Traditional	`Duration`	Duration of the insurance coverage of a given policy, in days
	`Insured.age`	Age of insured driver, in years
	`Insured.sex`	Sex of insured driver (Male/Female)
	`Car.age`	Age of vehicle, in years
	`Marital`	Marital status (Single/Married)
	`Car.use`	Use of vehicle: Private, Commute, Farmer, Commercial
	`Credit.score`	Credit score of insured driver
	`Region`	Type of region where driver lives: rural, urban
	`Annual.miles.drive`	Annual miles expected to be driven declared by driver
	`Years.noclaims`	Number of years without any claims
	`Territory`	Territorial location of vehicle
Telematics	`Annual.pct.driven`	Annualized percentage of time on the road
	`Total.miles.driven`	Total distance driven in miles
	`Pct.drive.xxx`	Percent of driving day xxx of the week: mon/tue/.../sun
	`Pct.drive.xhrs`	Percent vehicle driven within x hrs: 2hrs/3hrs/4hrs
	`Pct.drive.xxx`	Percent vehicle driven during xxx: wkday/wkend
	`Pct.drive.rushxx`	Percent of driving during xx rush hours: am/pm
	`Avgdays.week`	Mean number of days used per week
	`Accel.xxmiles`	Number of sudden acceleration 6/8/9/.../14 mph/s per 1000miles
	`Brake.xxmiles`	Number of sudden brakes 6/8/9/.../14 mph/s per 1000miles
	`Left.turn.intensityxx`	Number of left turn per 1000miles with intensity 08/09/10/11/12
	`Right.turn.intensityxx`	Number of right turn per 1000miles with intensity 08/09/10/11/12
Response	`NB_Claim`	Number of claims during observation
	`AMT_Claim`	Aggregated amount of claims during observation

UCONN

Data types of all 52 variables in the synthetic dataset

Category	Continuous/Integer	Percentage	Compositional
Marital	Duration	Annual.pct.driven	Pct.drive.mon
Insured.sex	Insured.age	Pct.drive.xhrs	Pct.drive.tue
Car.use	Car.age	Pct.drive.rushxx	.
Region	Credit.score		.
Territory	Annual.miles.drive		Pct.drive.sun
NB_Claim	Years.noclaims		Pct.drive.wkday
	Total.miles.driven		Pct.drive.wkend
	Avgdays.week		
	Accel.xxmiles		
	Brake.xxmiles		
	Left.turn.intensityxx		
	Right.turn.intensityxx		
	AMT_Claim		

The procedure: how?

- Data generation is a three-stage process.
  1. Simulating values for claims frequency as a multiple binary classification using feedforward neural network.
  2. Simulating values for claims severity as a regression using feedforward neural network with claims frequency feature.
  3. A synthetic portfolio of the space of feature variables is generated applying an extended `SMOTE` algorithm

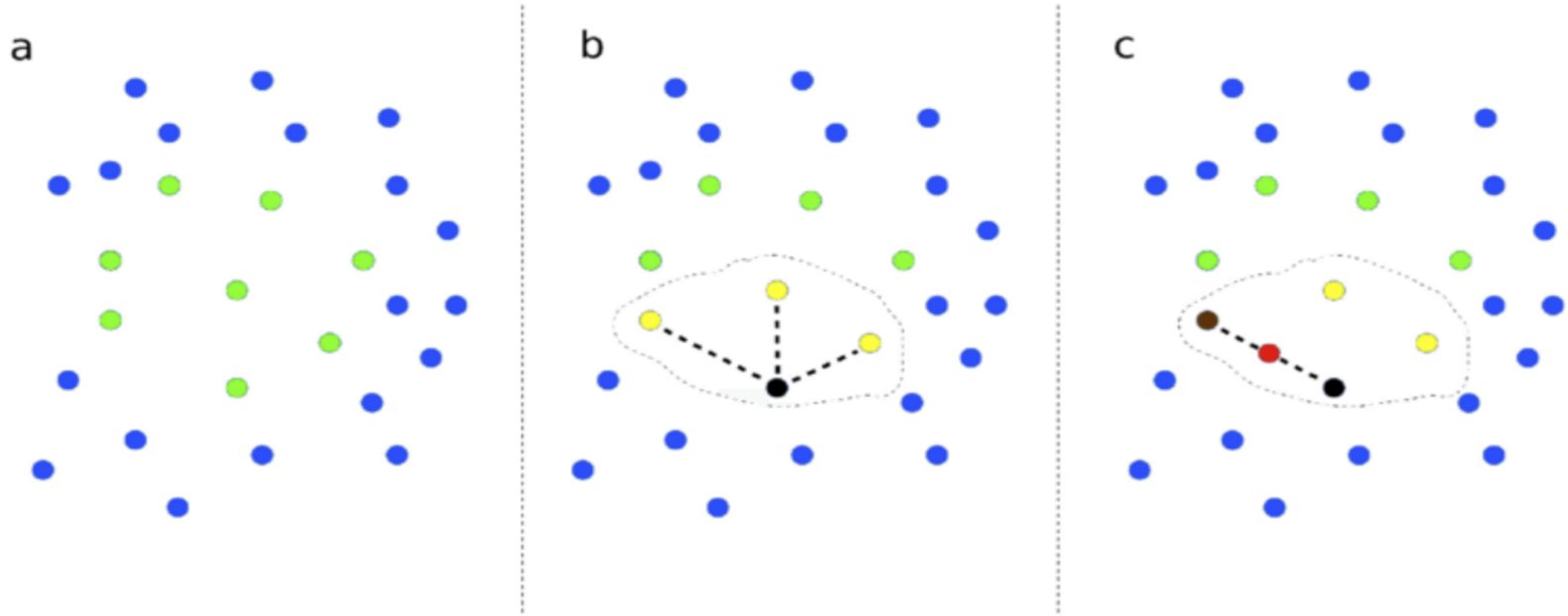
Why neural network?

- Popular algorithm known to give high accuracy of prediction.
- However, easy to overfit and difficult to know how it works (called black box model).
- From the perspective of synthesizing the real data, the downside of neural network is to our benefit.
  - Overfitting might help us to build simulation which very closely mimic the real response variables.
  - Hornik et al. (1989) proved that standard multi-layer feedforward networks are capable of approximating any measurable function, and thus are also referred to as universal approximators.
    - Lack of success in the applications usually due to inadequate learning, insufficient numbers of hidden units.
    - Tuning hyperparameters is essential in the process: Gaussian process (GP) is used.
    - Optimization is performed using `Adam` optimizer.

UCONN

Portfolio simulation: extended SMOTE

- The Synthetic Minority Oversampling Technique (`SMOTE`) is the process to create synthetic data points from minority class. [1]



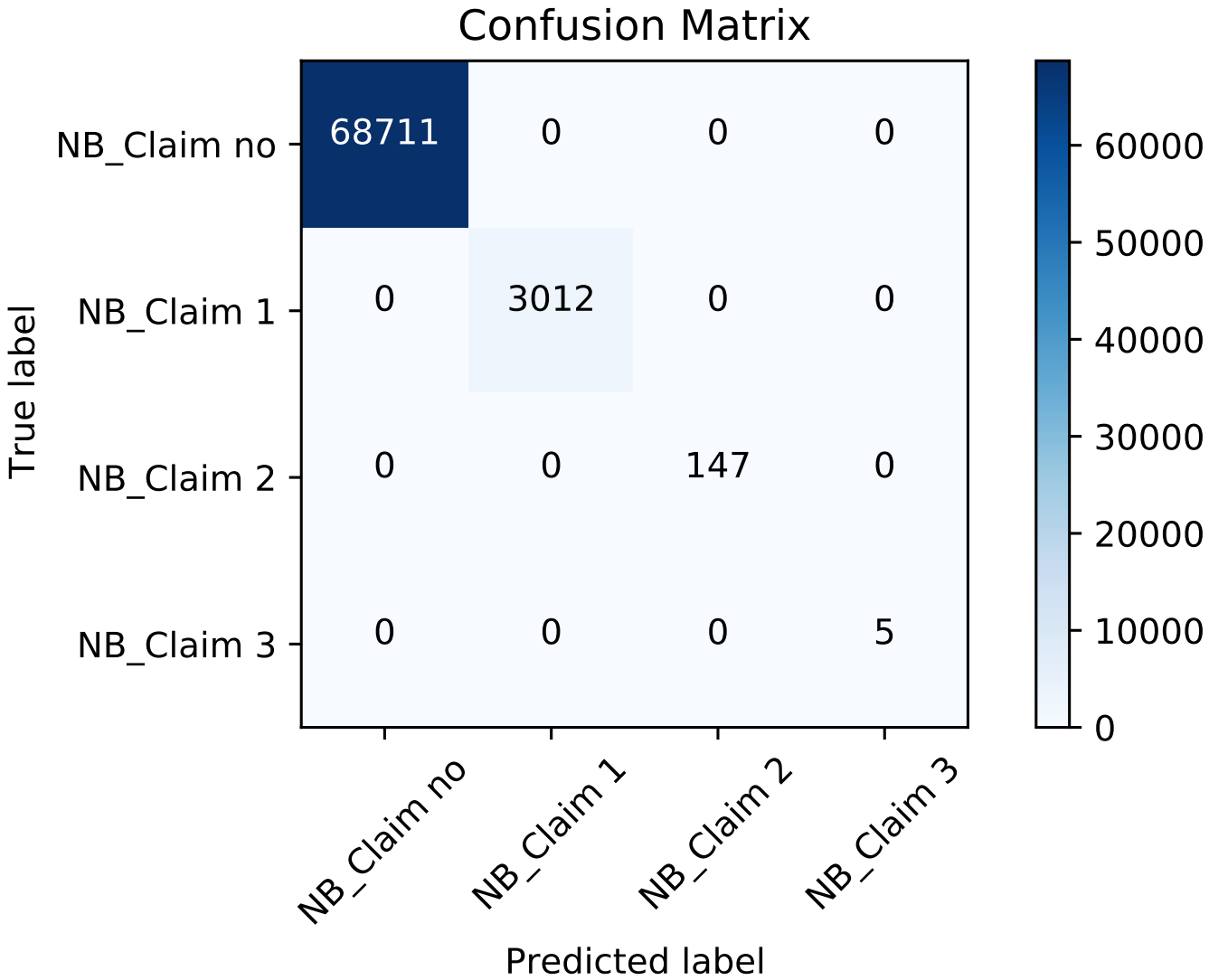
1. For each minority class instance, its K-nearest neighbors are obtained ($K = 3$ in the graph).
2. One of these K instances are randomly chosen to compute the new synthetic instances by linear interpolation.

[1] Image source: Schubach, et al. (2017)

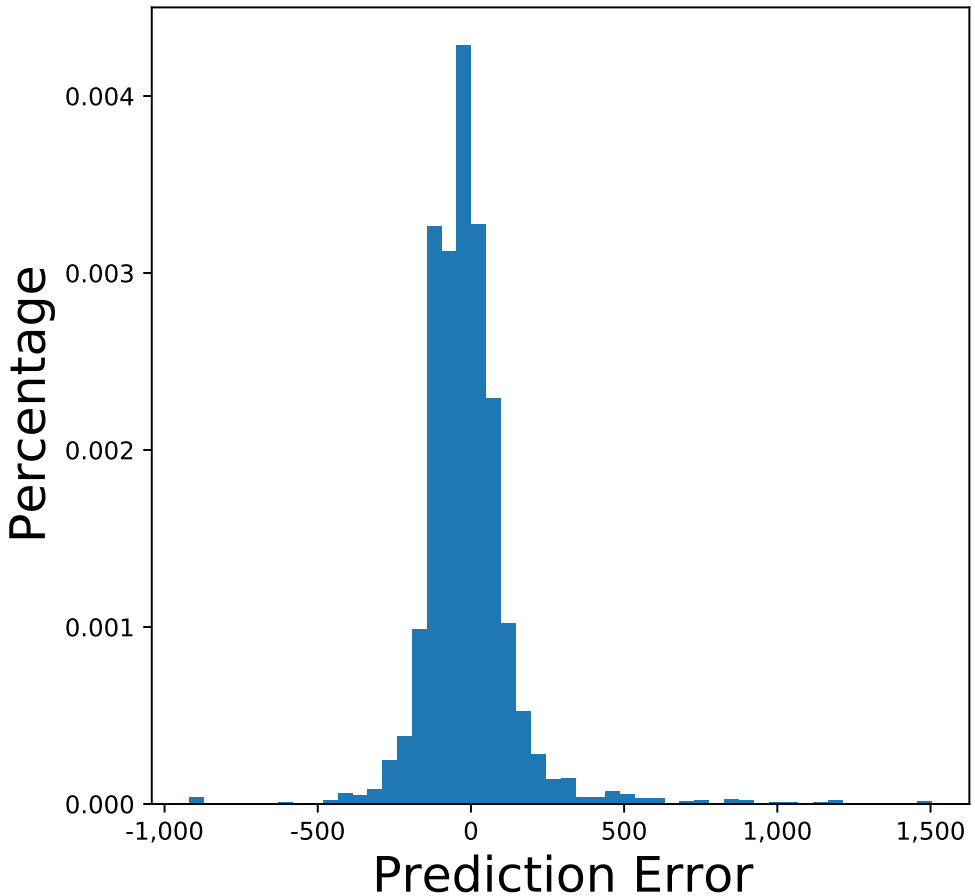
continued:

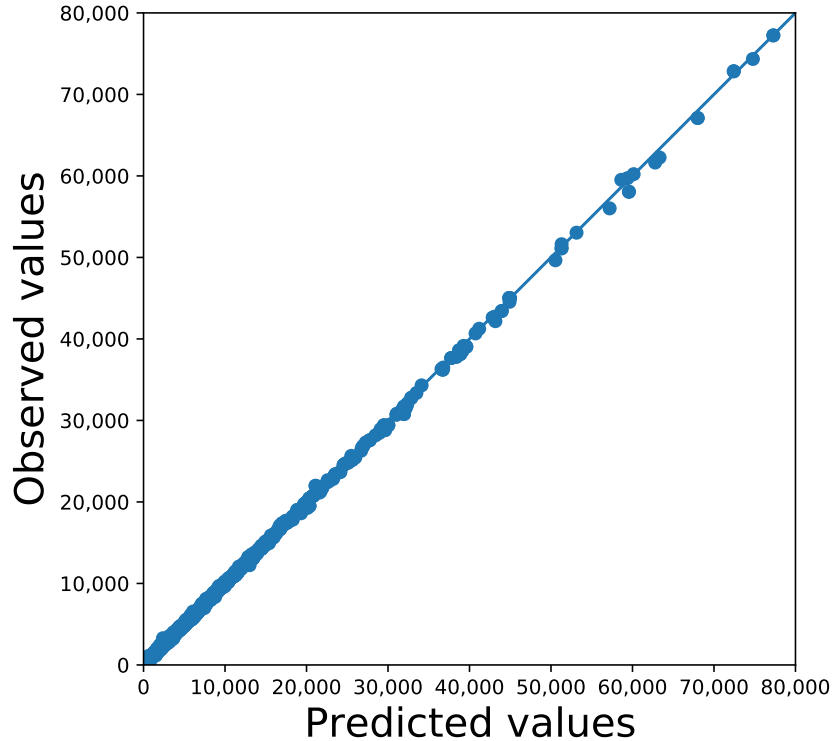
- Extend `SMOTE` to create synthetic instances from the entire dataset by changing one part of algorithm.
  - Linear interpolation $\rightarrow$ Interpolation from U-shape distribution.
- To preserve the characteristics of the original data distribution, especially in capturing outliers, we draw random numbers from U-shape distribution, not from uniform distribution, in the interpolation step.
  - New instance is closer to either chosen feature vector or the neighbor.
- Generate synthetic portfolio, X^s, consisting of 100,000 synthetic observations.
- This portfolio is used as the input of claim frequency simulation and claim severity simulation.

Claim frequency simulation - how well?



Claim amount simulation - how well?





Summary statistics: synthetic vs real

Synthetic	NB_Claim	Mean	Std Dev	Min	Q1	Median	Q3	Max
AMT_Claim	0	0	0	0	0	0	0	0
	1	4062	6767	0	670	2191	4776	138767
	2	8960	9554	0	2350	7034	11225	56780
	3	5437	2314	2896	3620	5372	5698	9743

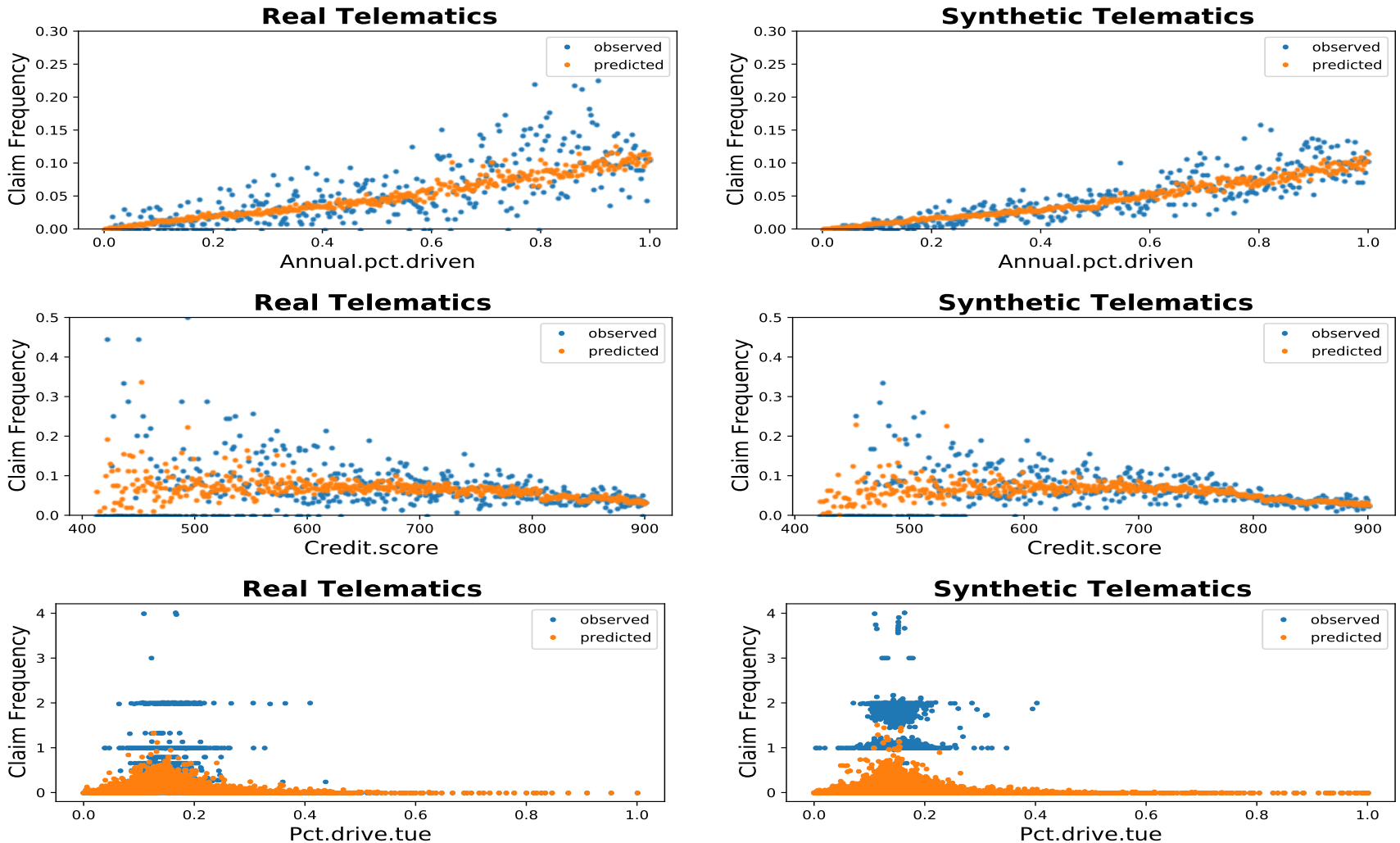
Real	NB_Claim	Mean	Std Dev	Min	Q1	Median	Q3	Max
AMT_Claim	0	0	0	0	0	0	0	0
	1	4646	8387	0	659	2238	5140	145153
	2	8643	10920	0	1739	5184	11082	62259
	3	5682	2079	3253	4540	5416	5773	9521

Comparison: Poisson and gamma regression

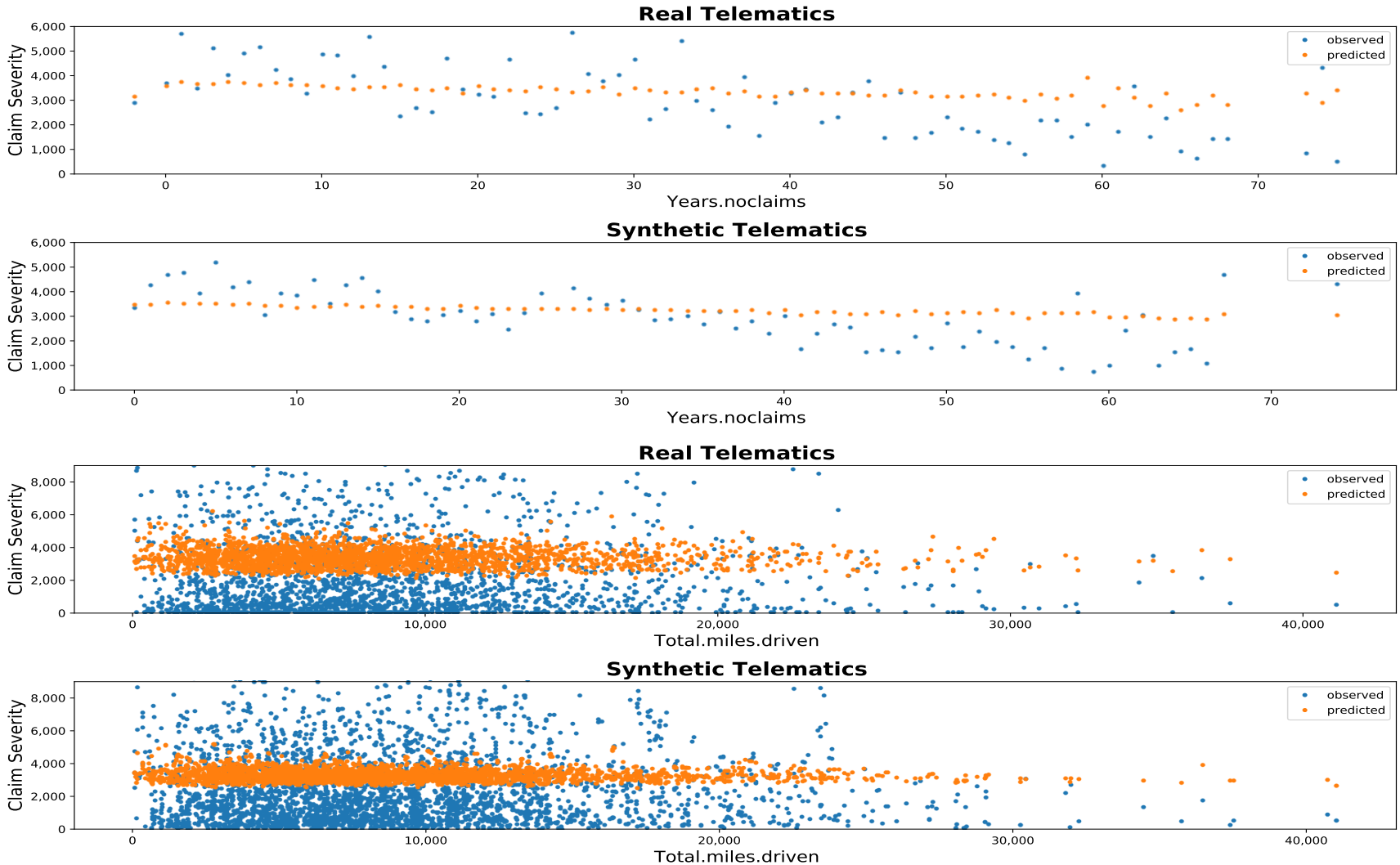
- The resulting data generation is evaluated by a comparison between the synthetic data and the real data.
  - Poisson and gamma regression models are fitted to the respective data.
- In the analysis, we predict the frequency $(\frac{\text{NB\_Claim}}{\text{Duration}})$ via Poisson regression and severity $(\frac{\text{AMT\_Claim}}{\text{NB\_Claim}}|\text{NB\_Claim} > 0)$ via Gamma.

$$\texttt{Premium} = \text{E}\left(\frac{\text{NB\_Claim}}{\text{Duration}}\right) \times \text{E}\left(\frac{\text{AMT\_Claim}}{\text{NB\_Claim}}|\text{NB\_Claim} > 0\right)$$

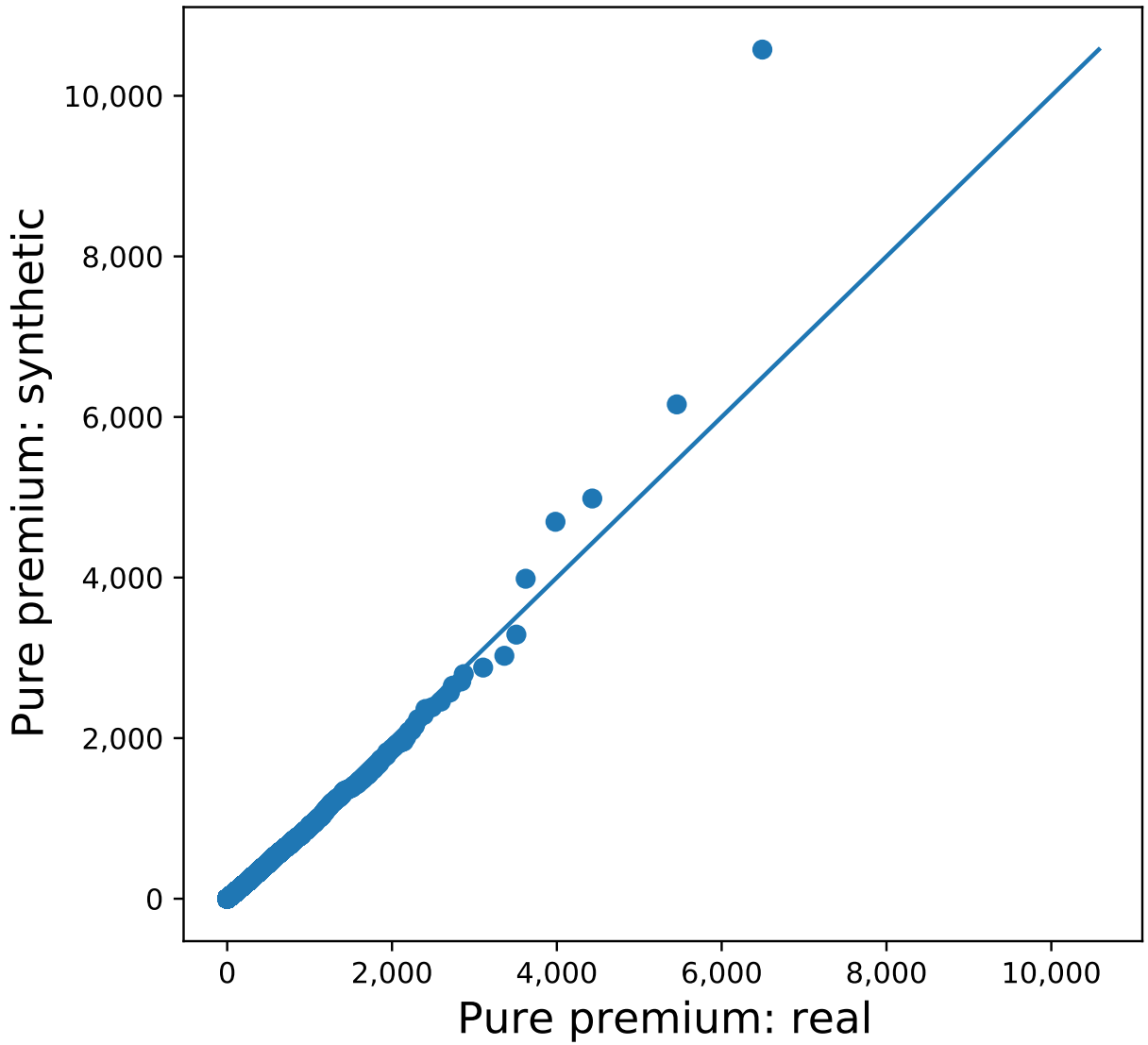
Left: Real, Right: Synthetic,
X-axis: Key Feature, Y-axis: Average Frequency



1st & 3rd: Real, 2nd & 4th: Synthetic, X-axis: Key Feature & Y-axis: Average Severity



QQ-plot of pure premium



To access dataset: where?

So, B., Boucher, J.-P., and Valdez, E.A. (2021). Synthetic Dataset Generation of Driver Telematics. *Risks*, 9(4), 58. Access here: `https://doi.org/10.3390/risks9040058`

Data available here:
`http://www2.math.uconn.edu/~valdez/data.html`

Selected references


Goodfellow, I., Bengio, Y., and Courville, A. (2016) *Deep Learning*. MIT Press.


Hastie, T. et al. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer: New York.


Chawla, N.V. et al. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 16: 321-357.


Dalkilic, T.E., Tank, F., and Kula, K.S. (2009). Neural networks approach for determining total claim amounts in insurance. *Insurance: Mathematics and Economics*. 45(2), 236-241.


Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L. and Sales, A.P. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*. 20, 108.


So, B., Boucher, J.P., and Valdez, E.A. (2020). Cost-sensitive Multi-class AdaBoost for Understanding Driving Behavior with Telematics. Available at arXiv: `https://arxiv.org/pdf/2007.03100.pdf`