Institute and Faculty of Actuaries

Alternative Data for GI pricing

Buu Truong and Mark Lee
Insight Risk Consulting

Contents

- Why this conversation now?
- What data is currently used?
- Data augmentation
- What is alternative data?
- What is the data business model?
- New data: Data Engineering
- New data: Data Science
- So what now?

Why this conversation now?

- Conversation has moved on from Big Data to more targeted uses cases of data.
- Insurance pricing should be data driven supplemented with judgement. There is a long way to go, particularly for commercial pricing.
- Alternative data could lead to:
  - Improved underwriting experience
  - Improved pricing refinement
  - Faster and more accurate claims settlement
- In the investment industry, alternative data is well established. Hundreds of millions (USD) spent a year on alternative data in the search for ‘alpha’.

What data is currently used?

- Personal lines: Proposal form data supplemented with…
  - Vehicle data
  - Personal data
  - Financial data such as credit scores
  - Property data (natural perils)
  - Property data (building specific)
- Commercial lines: Proposal form data supplemented with…
  - Property data (natural perils)
  - Some corporate and financial data
  - Limited use of specialist look-ups
  - Qualitative risk reports

Data augmentation

- Data augmentation is a grey area which starts at data backfill.
  - Example of backfill might be car details from DVLA for motor.
  - Example of augmentation might be flood score from proprietary provider for home.
  - The theoretical split between backfill and augmentation is likely to relate to how much the policyholder knows about what is happening.
- Advantages of data augmentation include:
  - For information known to the policyholder, saves policyholder time where it can be separately sourced / verified.
  - For information not known to the policyholder, adds new information that can support pricing.

What is alternative data?

Insight Risk Consulting

- Aggregator data
- Telematics data
- Home IoT device data
- Existing policy and claims data
- Third party company scores and review data
- Media / News data
- Risk evaluation data
- Companies house data
- Social network data
- Government data

Institute and Faculty of Actuaries

What is the data business model?

- Data providers
  - Unique data sets – from hardware or historical advantage
  - Scraped data sets – ubiquitous but not easily collated
  - Tidied up data sets – available but not always directly applicable
- Data pipelines
  - Data sets sold as a flat file or pay per click.
  - Preferred pipelines links to type of business as well as data use.
  - APIs as well as conventional feeds depending on velocity of data need.
- Difficulties may include IP issues where data ownership may be questioned or data ownership needs to be protected.

New data: Data Engineering

- Most technical challenges relate to live pricing environments where volume is high – such as current personal lines but in future likely for SME commercial.
- Two common challenges:
  - Ingesting large databases.
  - Use of API calls for live/on demand data.
- There is an increased need for actuaries and data scientists to work with developers to implement data collection, live pricing and data storage.
- There are different data challenges for low volume business though typically this might not called be called ‘data engineering’.

Data Engineering – large databases

- Postcode level data in the UK comes with 1.7 million rows. Property level data comes with ~30 million rows. This is beyond Excel.
- To match these data files with other data, either database queries (Access, SQL, NoSQL, etc) or programming tools like R or Python are needed.

Data Engineering - APIs

- Many suppliers of data will use RESTful API's
  - Advantages:
    - Supplier manages updates, live access to latest data.
    - Get only what you need – possibly a price per click, manageable data volumes.
- RESTful API's – essentially http requests, as for a webpage:
  - E.g.: https://dvlasearch.appspot.com/DvlaSearch?apikey=DvlaSearchDemoAccount&licencePlate=mt09nks

Data Engineering - APIs

- New skills – ability to programmatically run multiple web queries: R, Python, curl etc.
- Understanding how to parameterise the queries to get the appropriate data.
- Programmatically reading the responses – json/XML files are common – and loading the relevant fields to match to other data.
- Link data from multiple sources.

{"taxed":false,
"mot":true,
"dateOfFirstRegistration":"23 JULY 2009",
"yearOfManufacture":"2009",
"make":"VOLKSWAGEN",
"model":"TIGUAN SE TDI 4MOTION",
"fuelType":"DIESEL",
"sixMonthRate":"",
"twelveMonthRate":"",
"cylinderCapacity":"1968 cc",
"wheelPlan":"2-AXLE-RIGID BODY",
"revenueWeight":"Not available",
"taxDetails":"Tax due: 01 February 2019",
"taxStatus":"Not taxed",
"colour":"SILVER",
"typeApproval":"M1",
"co2Emissions":"167 g/km",
"motDetails":"Expires: 10 May 2019",
"numberOfDoors":5,
"vin":"XXXXXXXXXXXXXXXXX",
"transmission":"MANUAL"}

Data Engineering - APIs

- Response structure can be more complicated – an example json:


```
{
  "make": "VOLKSWAGEN",
  "model": "TIGUAN",
  "dateFirstUsed": "23 JULY 2009",
  "fuelType": "DIESEL",
  "colour": "SILVER",
  "engineSize": "1968",
  "registrationDate": "23 JULY 2009",
  "manufactureDate": "23 JULY 2009",
  "manufactureYear": "2009",
  "motTestReports": [
   {
    "testDate": "11 MAY 2018",
    "expiryDate": "10 MAY 2019",
    "testResult": "PASS",
    "odometerReading": 88237,
    "odometerUnit": "mi",
    "motTestNumber": 246230668405,
    "advisoryItems": [
      "Front Tyre worn close to the legal limit  Both (4.1.E.1)",
      "Rear  Both Tyre a have low cut on tread","Front Anti-roll bar linkage ball joint dust cover damaged, but preventing the ingress of dirt  Both (2.4.G.2)"
    ],
    "minorItems": [],
    "failureItems": []
   },
   {
    "testDate": "11 MAY 2018",
    "expiryDate": "",
    "testResult": "FAIL",
    "odometerReading": 88237,
    "odometerUnit": "mi",
    "motTestNumber": 436849190066,
    "advisoryItems": [
      "Front Tyre worn close to the legal limit  Both (4.1.E.1)",
      "Rear  Both Tyre a have low cut on tread",
      "Front Anti-roll bar linkage ball joint dust cover damaged, but preventing the ingress of dirt  Both (2.4.G.2)"
    ],
…
```

Data Engineering – missing data

- Data provided from API's is often "raw".
- Missing data is a common problem. Depending on the API, the responses to individual fields may be missing, the fields themselves may be missing.
- Different responses may have different combinations of fields.
- Need to handle exceptions – impute/ask/default/decline?
- Response time can be an issue – e.g., when supplying an aggregator.

{"taxed":false,
"mot":true,
"dateOfFirstRegistration":"23 JULY 2009",
"yearOfManufacture":"2009",
"make":"VOLKSWAGEN",
"model":"TIGUAN SE TDI 4MOTION",
"fuelType":"DIESEL",
"sixMonthRate":"",
"twelveMonthRate":"",
"cylinderCapacity":"1968 cc",
"wheelPlan":"2-AXLE-RIGID BODY",
"revenueWeight":"Not available",
"taxDetails":"Tax due: 01 February 2019",
"taxStatus":"Not taxed",
"colour":"SILVER",
"typeApproval":"M1",
"co2Emissions":"167 g/km",
"motDetails":"Expires: 10 May 2019",
"numberOfDoors":5,
"vin":"XXXXXXXXXXXXXXXXX",
"transmission":"MANUAL"}

Data Engineering – data storage

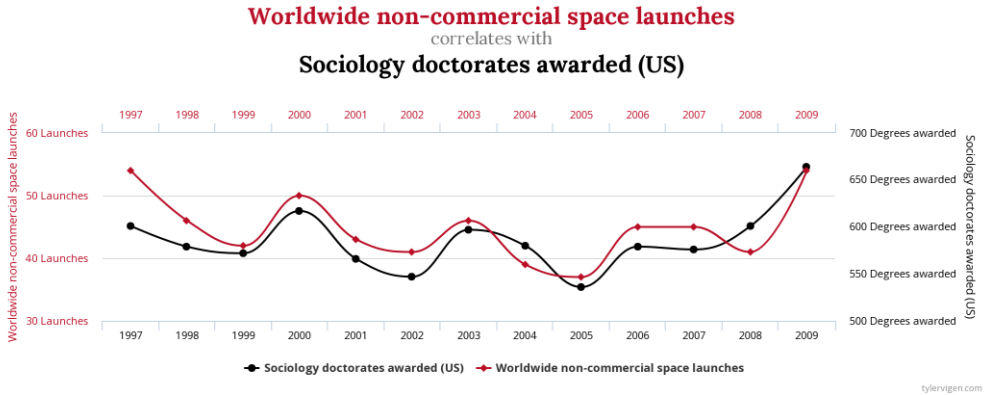
- Even if you are only using a couple of fields in the API response to price on, you might want to save the entire response in order to look for correlations in the future.
- API responses with many potential fields have to be stored in a data warehouse.
- For frequently refreshed data, the data may need to be collected or monitored by time.
- This can lead to “big data” – e.g. telematics raw data.

New data: Data Science

- So you now have lots of new data, so what? **Is it predictive?**
- Prove value by historical claims analysis
  - Can you backfill the data to match your back-book?
  - Potentially many gaps – can you impute?
  - If not, can you make a case to collect for future analysis?
- Short term vs long term value
  - First movers may get significant advantage, but value may change once market uses new data as standard
  - On the other hand, if you don't get data that becomes market standard, a high risk of being selected against.
- Cost of data vs Value from data – what is the appropriate ratio?

Data Science – correlations

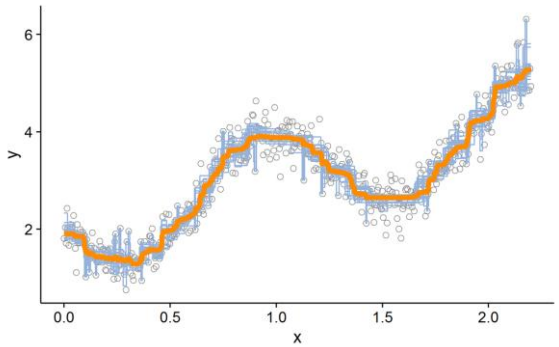
- A statistical exercise of finding a signal
  - With multiple new fields, the chance of a variable looking predictive by chance is much increased.
  - Look at correlations, but bear in mind that correlation does not imply causation. Understanding how predictions generalise to unseen data is crucial – use test sets or cross-validation.

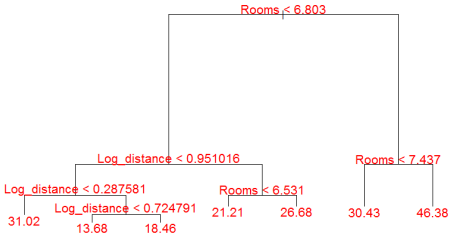


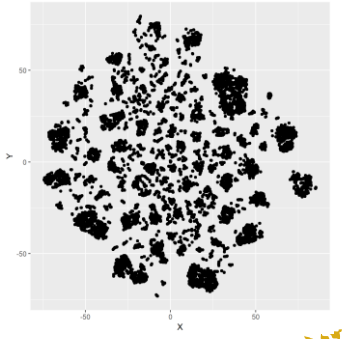
www.tylervigen.com/spurious-correlations (CC licence)

Data Science - techniques

- Use new predictive models – e.g., gradient boosted models, random forests, neural networks, support vector machines. R or python are useful here.
- Use unsupervised analysis (clustering, dimensionality reduction) to look for interactions affecting just small proportion of data, or complex interactions.

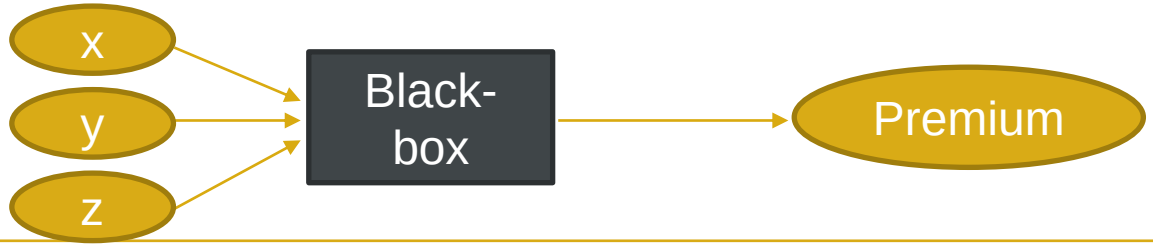






Data Science - transparency

- Advanced and flexible models can be difficult to interpret – black-box like.
  - Can you explain to other stakeholders?
    - data visualisation – one-way plots,
    - developing approximate but transparent models (e.g., GLMs) to explain trends,
    - communication of test results.
  - Are you certain that the routine is not discriminating on, e.g., Gender or Race? Can you demonstrate this to a regulator?



So what now…

- We should think more about data engineering within our underwriting and pricing frameworks.
- In a blog last year we talked about data actuaries (and finance actuaries). I think this is an increasing trend.
- Seek out opportunities within your firms to get involved with proof of concept work.
- Our previous talks on parameter error and increasing statistical robustness in London Market pricing (at GIRO and LMAG) align to data adding value.
- More data means more modelling, and more actuarially focused pricing. This is good news for the profession!

Institute and Faculty of Actuaries

Alternative Data for GI pricing

Buu Truong and Mark Lee
Insight Risk Consulting