Penalised Regression
Tony Ward
tony@statcore.co.uk
Will a Robot Take your job in the next 20 years?

Actuaries, economists and statisticians

Likelihood of automation?
It's quite unlikely (15%)

How this compares with other jobs:
263rd of 366
Will a Data Scientist take your job in the next 20 years?
Today we will cover:

- How to improve the accuracy of predictive models using unstructured data

- How to replace an entire risk pricing modelling team with one Data Scientist
How to improve the accuracy of predictive models using unstructured data
Motivating Example: Salary Prediction

- Web Scraped 28,000 Jobs from [www.theactuaryjobs.com](http://www.theactuaryjobs.com)
- Want to predict the salary based on the data provided in the advert.
- Data is both structured (sector, experience) and unstructured (job title, further info)
Global insurance company with Lloyds syndicate is looking to hire a pricing actuary reporting directly into the Group pricing actuary. You will have experience of pricing and be looking to lead a small team based in London. The business covers a variety of lines of business including marine, aviation, energy, property and liability. You will work closely with the underwriters, present to senior management and underwriters and own and manage rate adequacy and rate movement MI for international specialty lines of business. Above and beyond your pricing skills you must be able to demonstrate leadership qualities and have excellent communication skills.
Job 2

Our client, an investment consultancy, is looking to hire a research analyst. You will work on investment manager research projects. Working closely with the investment consulting team, you will conduct research on multi-asset fund managers and equity fund managers. Candidates are expected to have experience in researching investment managers and an interest in financial markets and funds. You must be a good team player with good communication skills and be able to present to the investment committee. Our client offers study support (e.g. CFA or actuarial qualifications)

<table>
<thead>
<tr>
<th>Recruiter</th>
<th>[Blacked Out]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>London</td>
</tr>
<tr>
<td>Salary</td>
<td>£35k - 40k + bonus + benefits</td>
</tr>
<tr>
<td>Posted</td>
<td>18 May 2015</td>
</tr>
<tr>
<td>Closes</td>
<td>18 Jun 2015</td>
</tr>
<tr>
<td>Ref</td>
<td>Research Analyst</td>
</tr>
<tr>
<td>Sector</td>
<td>Investment, Management consultancy, Pensions</td>
</tr>
<tr>
<td>Experience</td>
<td>Graduate/post-graduate, Nearly qual (11+ exams), Other, Part qual (1-10 exams)</td>
</tr>
<tr>
<td>Contract Type</td>
<td>Permanent</td>
</tr>
<tr>
<td>Hours</td>
<td>Full Time</td>
</tr>
</tbody>
</table>
Some roles pay less than other!

They want volunteers?

### Pricing Actuary - Retail

- **Recruiter**: [Redacted]
- **Location**: London
- **Salary**: £0 per annum
- **Posted**: 23 Dec 2013
- **Closes**: 20 Jan 2014
- **Ref**: JK10399
- **Contact**: [Redacted]
- **Sector**: General insurance
- **Experience**: Nearly qual (11+ exams), Qualified
- **Contract Type**: Permanent
- **Hours**: Full Time
Setting up the Problem

- **TASK:** To predict the probability of a job paying more than £70,000

- **INPUT VARIABLES**
  - Sector: General Insurance, Pensions, ...
  - Experience: Graduate/post-graduate, Part qual (1-10 exams), ...
  - City
  - County
  - Job title e.g. Head of International Pricing
  - Further Info e.g. Global insurance company with Lloyds syndicate is looking to hire a pricing actuary reporting directly...
How to handle unstructured data

- Turn the unstructured data into a structured tabular form (a Document Term Matrix).

<table>
<thead>
<tr>
<th>Job Title</th>
<th>Analyst</th>
<th>Actuari</th>
<th>Head</th>
<th>International</th>
<th>Pricing</th>
<th>Research</th>
<th>Retail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head of International Pricing</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research Analyst</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Pricing Actuary - Retail</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

- The presence/absence of each word are the potential predictors in our model
- Do this for both “Job Title” and “Further Info” text items
- Now our modelling dataset is very wide!!
Modelling Dataset

<table>
<thead>
<tr>
<th>Rating Factor</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience</td>
<td>7</td>
</tr>
<tr>
<td>Sector</td>
<td>14</td>
</tr>
<tr>
<td>Country</td>
<td>89</td>
</tr>
<tr>
<td>City</td>
<td>99</td>
</tr>
<tr>
<td>Job Title</td>
<td>189</td>
</tr>
<tr>
<td>Further Info</td>
<td>691</td>
</tr>
</tbody>
</table>

- Over 1000 different categorical levels to consider
- Fit a GLM with binomial error term (Logistic Regression).
Statistical Modelling - process

- One way tables
- Correlated Variables
- Forward / Backward Selection
- Statistical Tests of Coefficients
- Likelihood Ratio Test of Nested Models
- Interaction with Time
- Interaction with Random Factors
- Check Residuals
Time consuming!

There HAS to be a better way!
Penalised Regression

- Penalised Regression refers to the family of methods for building Generalised Linear Models using Regularisation.
- Also known as “Shrinkage Methods” since they shrink model coefficients towards zero.
- Methods include Lasso Regression, Ridge Regression and Elastic Net.
- Widely used in the machine learning community
- Used in insurance?
Penalised Regression

• Same model form as traditional GLM.
  – Error Term (Poisson, Gamma, Binomial)
  – link function (log, logit)
  – Offsets
  – Relativities

• Key difference is how we build the model

• Same process as with other machine learning algorithms: k-fold cross validation to optimise the model hyper-parameters
Lasso Example: Salary Prediction

Model Parameters: Lasso Regression vs. GLM

- **Linear Predictor**
  - Gradient: -2 to 2
  - Values: -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2

- **No. Job Adverts**
  - Values: 0, 1500, 3000, 4500, 6000, 7500, 9000, 10500, 12000

### Job Adverts (Split by Experience)
- **Categories**: Graduate/Post Grad, Part Qualified, Nearly Qualified, Qualified, Qualified (Exec), Other, Banking & Finance, General Insurance, Health, Hedge Funds, Investment, IT, Life Insurance, Management Consultancy, Other, Pensions, Reinsurance, Risk Management, Solvency II, Systems

### Job Adverts (Split by Sector)
- **Categories**: Job Adverts (Split by Experience) - Linear Predictor vs. No. Job Adverts

---

**Lasso Regression** vs. **GLM**

- **Graph Comparison**

---

**Colour palette for PowerPoint presentations**
- **Dark Blue**: R17, G52, B88
- **Gold**: R217, G171, B22
- **Mid Blue**: R64, G150, B184
- **Secondary colour palette**
  - **Light Grey**: R220, G221, B217
  - **Pea Green**: R121, G163, B42
  - **Forest Green**: R0, G132, B82
  - **Bottle Green**: R17, G179, B162
  - **Cyan**: R0, G156, B200
  - **Light Blue**: R124, G179, B225
  - **Violet**: R128, G118, B207
  - **Purple**: R143, G70, B147
  - **Fuscia**: R233, G69, B140
  - **Red**: R200, G30, B69
  - **Orange**: R238, G116, B29
## Candidate Models

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Factors Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>Sector</td>
</tr>
<tr>
<td>E</td>
<td>Experience</td>
</tr>
<tr>
<td>ES</td>
<td>Experience, Sector</td>
</tr>
<tr>
<td>ESCI</td>
<td>Experience, Sector, City</td>
</tr>
<tr>
<td>ESCICO</td>
<td>Experience, Sector, City, Country</td>
</tr>
<tr>
<td>ESCICOJ</td>
<td>Experience, Sector, City, Country, Job Title</td>
</tr>
<tr>
<td>ESCICOF</td>
<td>Experience, Sector, City, Country, Further Info</td>
</tr>
<tr>
<td>ESCICOJF</td>
<td>Experience, Sector, City, Country, Job Title, Further Info</td>
</tr>
</tbody>
</table>
Results

Model ESCICOJ which included “Job Title” performed the best.

Jobs containing “chief” “head” “expert” “director” got higher predictions.

Jobs containing “junior” “trainee” “student” received lower predictions.

Adding “Further Info” damaged performance.
Learnings

• Unstructured data can boost predictive performance
• Penalised Regression isn’t a silver bullet – throwing all your data at it doesn’t work.
• I’m glad I work in General Insurance and not Pensions!
How does Lasso Regression work?
Traditional GLM: Maximum Likelihood

- The GLM relativities are found by maximum likelihood.
- In practice we minimise the negative of the log likelihood

$$\min_{\beta_0, \beta}
\left\{-l(\beta | X, Y)\right\}$$

- $\beta_0$: Intercept
- $\beta$: Relativities
- $l(\beta | X, Y)$: Log Likelihood
Lasso Regression

- Lasso Regression fits the GLM subject to the following constraint

\[ \sum_{j=1}^{p} |\beta_j| \leq t \]

- Models are estimated by penalised maximum likelihood

\[ \min_{\beta_0, \beta} \left( -\frac{1}{N} l(\beta | X, Y) + \lambda \sum_{i=1}^{N} |\beta_i| \right) \text{ lihood} \]
Penalty Term Lambda

• We need to find the best $\lambda$ to use.
  
  – Big $\lambda$ = Strong Penalty $\gg$ Most coefficients set to zero
  – Small $\lambda$ = Weak Penalty $\gg$ Approaching Traditional GLM

• To select $\lambda$ we split the data into training and test. Using the training data we test a range of $\lambda$ values using 10-Fold Cross Validation.

• Lambda is chosen to minimise the cross validation error.
Fit a Lasso Regression

Given Lambda

Error 1
Error 2
Error 3
Error 4
Error 5
Error 6
Error 7
Error 8
Error 9
Error 10

Calculate Average Error

Choose lambda to minimise average error
Small Lambda / Weak Penalty

Coefficients enter model

Big Lambda / Strong Penalty

At this penalty we have 16 coefficients in the model – the rest are zero
How to replace an entire risk pricing modelling team with one Data Scientist
Statistical Modelling:

- Three or four analysts per product.
- Each person can only build one model at once

**Home Risk Pricing Team**
- FLOOD
- FIRE
- THEFT
- AD

**Motor Risk Pricing Team**
- TPD
- TPI
- AD
- THEFT

- The only way to speed this up is to hire more staff
- This doesn’t remove the endless debate around borderline decisions.

*You are the bottleneck!*
• Model building is automated therefore many models can be built in parallel
• No more debate about borderline decisions – remember the hedge fund example
• Quicker Speed to Market
• Models can be continuously updated as new data arrives.
• **Now the bottleneck is IT!**
Success criteria

- Speed to Market
- Interpretation
- Implementation
- Accuracy

Challenging!
Challenges

• Throwing all data at Penalised Regression doesn’t work.
• Data sets are big!
• Numerical Inputs
• Interactions
Advantages

- How much data should we include?
- 2 Accident Years?
- 3 Accident Years?
- Now we can make this decision scientifically – test them all!
My Solution – a machine learning pipeline

- Training Data
  - Algorithm 1
  - Algorithm 2
  - Algorithm 3
- Processed Data
- Validation Data
- Penalised Regression
- Model Output
Success criteria

- Speed to Market
- Interpretation
- Implementation
- Accuracy
Summary

• How unstructured data can improve models
• Introduction to Lasso Regression
• How to replace a risk pricing team with one Data Scientist
Thank you!

Tony Ward

www.statcore.co.uk

tony@statcore.co.uk
Appendix: ROC chart

Hold Out Performance: ROC

Area under ROC Curve

86% 87% 88% 89% 90% 91% 92% 93%
Appendix: Ridge Regression

- Ridge Regression fits the GLM subject to the following constraint

\[ \sum_{j=1}^{p} \beta_j^2 \leq t \]

- Models are \( \min_{\beta_0, \beta} \left( -\frac{1}{N} l(\beta | X, Y) + \lambda \sum_{i=1}^{N} \frac{\beta_i^2}{2} \right) \)
Strong Penalty > All coefficients are zero

Weak Penalty > All coefficients non-zero

All coefficients quickly enter the model
Lasso vs Ridge

Alpha=1: Lasso Regression

Alpha=0: Ridge Regression
Appendix: Lasso vs Ridge – Part 1

• To understand why the lasso shrinks coefficients to zero and the ridge does not consider the following. The chart depicts the lasso (left) and ridge regression (right) when there are only two parameters.
Appendix: Elastic Net

- The elastic-net penalty is controlled by $\alpha$, and bridges the gap between Lasso ($\alpha=1$) and Ridge ($\alpha=0$). The tuning parameter $\lambda$ controls the overall strength of the penalty.

- Models are estimated by penalised maximum likelihood

$$
\min_{\beta_0, \beta} \left( -\frac{1}{N} l(\beta|X, Y) + \lambda \left( (1 - \alpha) \sum_{i=1}^{N} \frac{\beta_i^2}{2} + \alpha \sum_{i=1}^{N} |\beta_i| \right) \right)
$$
Appendix: Lasso v Ridge v Elastic Net