Predicting Readmission of Diabetic Patients

Michael Gat

General Assembly Santa Monica, DSI Summer 2016

@michaelgat

http://github.com/michaelgat/DSI_Classwork/Capstone/DSI_Capstone_Final.pdf

Predicting Readmission of Diabetic Patients

- Problem statement
- Data
- Approach
    - Unique challenges
- Results
    - Good
    - Bad
    - Ugly
    - Top features
- Lessons learned
- Acknowledgements

The Problem

- Diabetes
  - Direct cause of ~9% of U.S. healthcare costs
  - Affects ~10% of the population
- Hospital Readmissions (patient has to come back)
  - < 30 day readmission rate is a key measure of quality of care
  - A big driver of costs, over $20b for Medicare program alone
  - There are significant penalties for high readmission rates
  - ***Often preventable!***
- Develop model to predict patients most likely to be readmitted.

Where data science fits in

- Preventing re-admission of patients has been a major focus
- Predictive analytics and machine learning have already had a major impact
  - Public models like LACE exist and are used widely
  - Health care organizations are investing in custom models to address specific concerns: http://www.healthcareitnews.com/blog/predictive-analytics-drive-down-hospital-readmissions
- More and better data is becoming available
- More and better tools are being deployed
  - Text analysis
  - Radiology/imaging analysis
  - Better algorithms for structured data

The Dataset

- 100,000 records of patients with a diabetic condition
- 62 useful features
- Classified into readmit/non-readmit
- Concerns
    - Lack any information that could compromise confidentiality
        - Location
        - Physician notes
        - Most personal characteristics
    - Many features are extremely sparsely populated
- Limitation
    - No cost data, so will be difficult to determine what measures to focus on.

The Dataset: Key variables

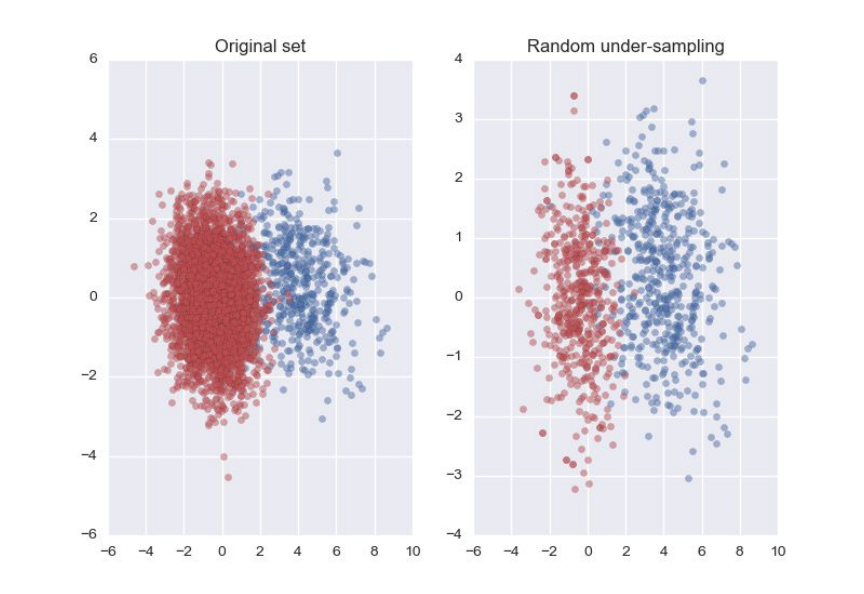
< Show plots of key variable value distributions in Capstone_1A notebook >

Approach

- Consult research literature
- Clean/reorganize data
- Develop simple baseline model
  - Subset of features
  - Simple feature selection
  - Test different classifiers
- Improve model in stages
  - Optimize feature selection
  - Optimize classifier
  - Over/under sampling of data
- Finalize model
  - Include all 62 available features.

Challenges: Sampling test data

- Over/under sampling to address an imbalanced dataset
- Chose under-sampling of ‘0’ (non-readmit) records
- Started writing my own routines to do this.
- Found (naturally) there is a package out there that will do it for me.
- Used Random Under Sampling (The simplest approach)
- Was critical to developing the model.
- http://contrib.scikit-learn.org/imbalanced-learn/auto_examples/index.html



Challenges: Feature selection

- 62 available features (potentially more!)
  - Many of these are related or associated with a common condition
- Identified Chi Squared as a likely feature selection mechanism
  - Cited frequently in related studies
  - A test of feature independence
- Widely used in health care
- Supported in scikit-learn
- Handles sparse data very well
  - Also used in some NLP situations for this reason
- Achieved better results than other tests
- https://en.wikipedia.org/wiki/Chi-squared_test

Results: Phase 1

- Selected 6 features using Chi Squared
- Best result with Naive Bayes:

	Non-Readmit (0)	Readmit (1)
Non-Readmit (0)	64797	2999
Readmit (1)	7421	1108

- Precision: 12.99%
- Specificity: 95.58%
- AUC: 0.5428
- This sucks! Maybe that river guide job in New Zealand wouldn't be a bad idea?

Results: Phase 2 - Better feature selection

- Best result with 8 features selected, using Chi Squared
- Best result with Naive Bayes:

	Non-Readmit (0)	Readmit (1)
Non-Readmit (0)	64285	3590
Readmit (1)	7254	1196

- Precision: 14.15%
- Specificity: 94.71%
- AUC: 0.5443
- This still sucks! Maybe I should find another project?

Results: Phase 3 - Add undersampling

- Best result with 8 features selected, using Chi Squared
- Best result with Naive Bayes:

	Non-Readmit (0)	Readmit (1)
Non-Readmit (0)	58383	9504
Readmit (1)	5988	2450

- Precision: 29.04%
- Specificity: 86.00%
- AUC: 0.5751
- Getting better. A doctor I know said it’s almost useful.

Results: Phase 4 - Use full feature set

- Best result with 14 features selected, using Chi Squared
- Best result with Naive Bayes:

	Non-Readmit (0)	Readmit (1)
Non-Readmit (0)	54810	13035
Readmit (1)	5557	2923

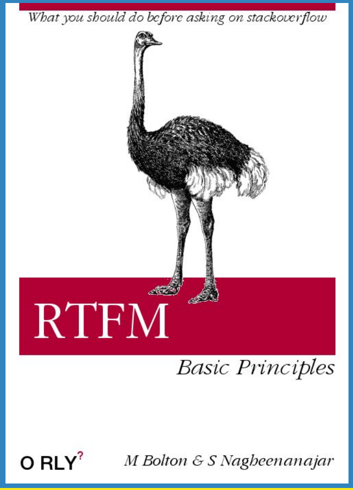
- Precision: 34.47%
- Specificity: 80.79%
- AUC: 0.5762
- Feeling OK. Approaches results I’ve seen in peer-reviewed research

Results: Top features

1. Discharge Disposition: Anything but “home”
2. Days in hospital: More is worse
3. Number of lab procedures
4. Number of medications given
5. Number of outpatient visits in prior year
6. Number of emergency visits in prior year
7. Number of inpatient visits in prior year
8. Number of distinct diagnoses
9. Received metformin (Glucophage)
10. Received insulin
11. IDC-9 diagnosis: 428 (Heart Failure)
12. IDC-9 diagnosis: 401 (Hypertension)
13. IDC-9 diagnosis: 403 (Hypertensive CRF)
14. IDC-9 diagnosis: 786 (Symptoms involving respiratory system and other chest symptoms)

Lessons learned

- In the real world, getting results is hard; small improvements take a lot of work
- Feature selection is huge when dealing with complex data
- Dealing with unbalanced data in an interesting wrinkle
- Not everything visualizes well
- Don’t reinvent the wheel; remember to use the basic tools we’ve all got



Acknowledgements

- Dataset: https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008
- Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records http://www.hindawi.com/journals/bmri/2014/781670/
- Predictive risk modelling for early hospital readmission of patients with diabetes in India http://link.springer.com/article/10.1007/s13410-016-0511-8
- Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure Patients https://cwds.uw.edu/sites/default/files/publications/Big%20Data%20Solutions%20for%20Predicting%20Risk-of-Readmission%20for%20Congestive%20Heart%20Failure.pdf
- @thepracticaldev (O'RLY parodies)

- Everybody in the class
- John, Pauline, Mike

Questions?

http://github.com/michaelgat/DSI_Classwork/DSI_Capstone_Final.pdf

@michaelgat