B.A.U. for actuaries:
Big data, Analytics & Unstructured data

Mudit Gupta
Big Data Working Party

Ling Yit Wong
Holmusk

SAS Health and Retirement Conference 2016

Big Data Working Party

Our mission

To explore big data, analytics and unstructured data solutions for use in Asia

To understand what actuaries need to do to have the right skillset such that we are equipped to be the forefront of demand for such work

To promote the use of these methodologies by actuaries via open knowledge-sharing

Who are we?

The working party is made up of actuaries from life insurance, general insurance and consulting backgrounds as well as data scientists in Singapore, Malaysia and Hong Kong

Current members

Mudit Gupta (Chair), Alvin Choong, Colin Priest, David Menezes, Frank Devlin, Frankie Chan, Huang Guoyu, Jerry Yu, Kate Chen, Kok Fai Wai, Paul Wang

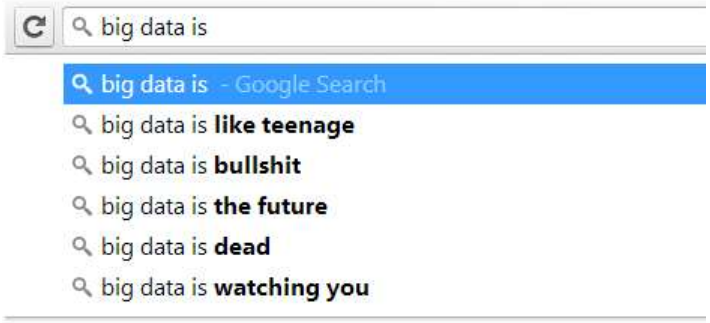
Big Data Working Party

Key achievements

- Developed **case studies** to demonstrate machine learning techniques
- **Presentations** and published **articles** at conferences and CPD events afternoon forums in Singapore, Hong Kong and Malaysia
- Organized hands-on **workshops on machine learning and text mining** using R with participants from all around Asia
- Supporting regional societies by sharing knowledge and in organizing of CPD events and workshops (recently in Thailand and Malaysia)
- Asia Actuarial Analytics Challenge 2016

Where to from here?

- Developing practical case studies which actuaries may apply in business situations
- We are also looking out to organizations wishing to collaborate with us on research and development in practical applications of data analytics

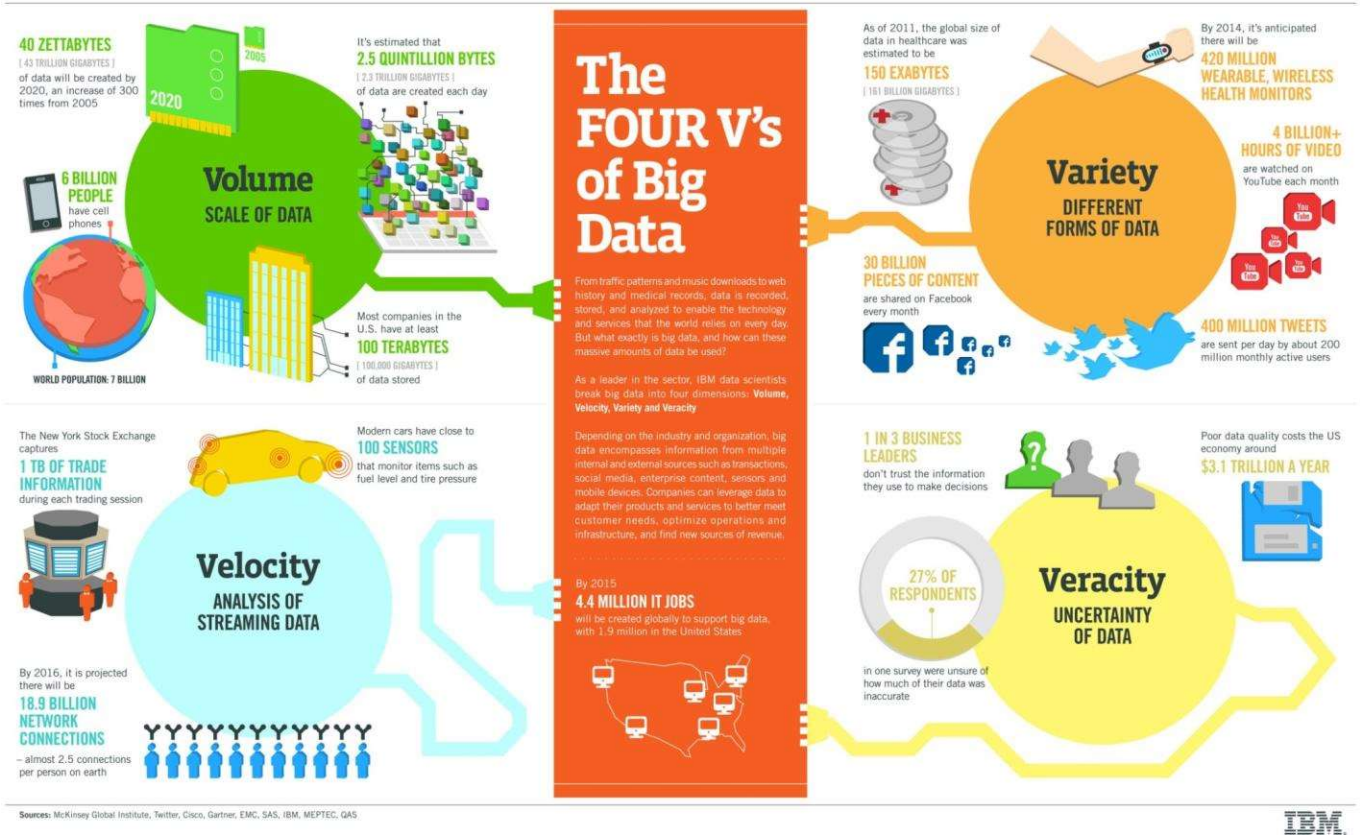


Introduction to Big data

Part I

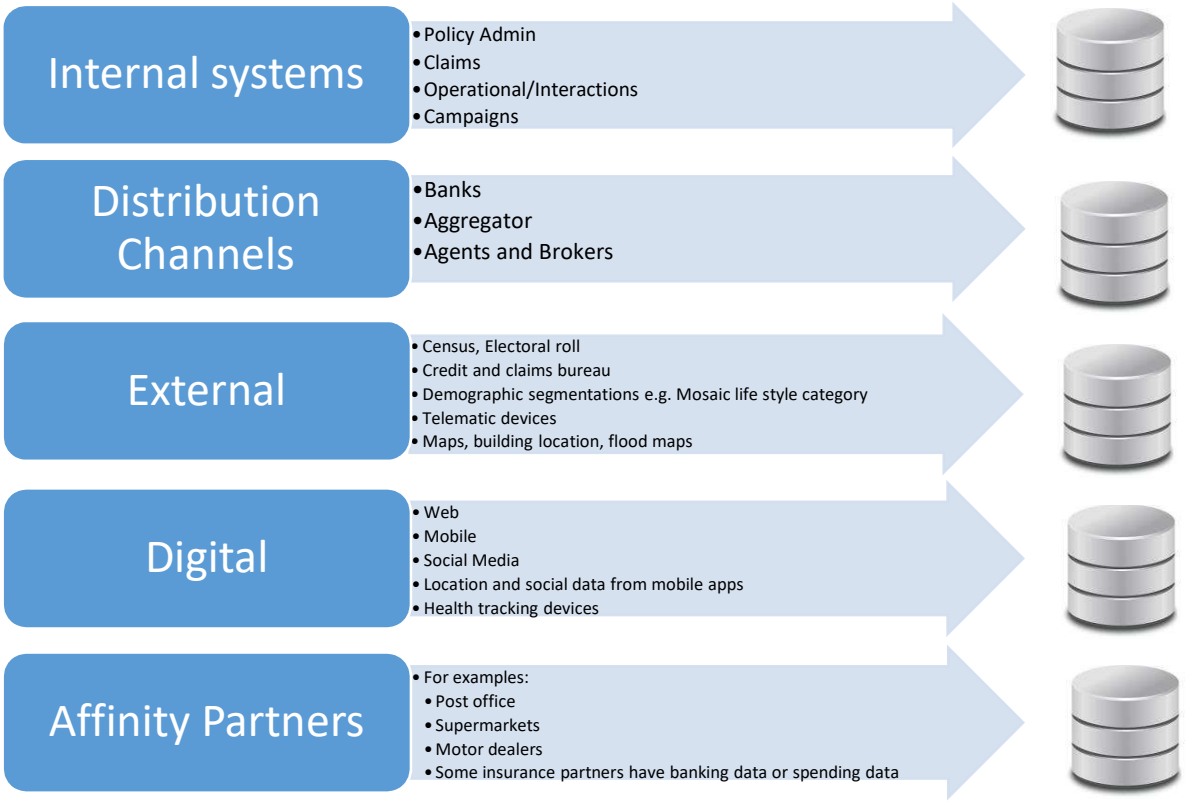
What is big data?

- Often used to describe large volume of data being collected by organizations
- Lack of structure



Infographic from http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Where does data come from?



Applications of Big Data in Insurance

Claims	Product & Pricing	Customer	Marketing & Distribution
•Fraud detection •Case Estimation •Claim handling •Loss adjustment	•Telematics •Wearable •Connected homes •Underwriting Acceptance •Granular Pricing •Price Comparison Websites (PCWs)	•Price Sensitivity •Up-sell / Cross-sell •Take up and Churn •Personalised targeting and messages •Customer satisfaction	•Agency Scoring •Orphan policies agent matching •Media effectiveness •Competitor analysis •Location-based targeting •Sentiment analysis

HR departments using big data – article in Financial Times 9 July 2014:

"Employees who are members of one or two social networks were found to stay in their job for longer than those who belonged to four or more social networks"

Potential for behavioral change

Privacy and ethical considerations

Predictive Underwriting using External Information

Life insurance in Thailand

- Swiss Re built a predictive underwriting model for a major Thailand life insurance company together with a local bank whereby the model predicts using banking information a prospective customer's chance of being a good/bad risk.
- Using the model they are able to select customers with good predicted underwriting risk, and offer them insurance without any additional underwriting.
- A Swiss Re blog on data analytics describes some valuable sources of data:
    - **Banks** have heavily invested in data and are exceptionally well placed to take advantage of their data
    - **Third party data sources** can have very strong predictive power in some markets
    - **Loyalty card / supermarket data** is frequently as strong – if not stronger – than banking data. The challenge is persuading these providers to extract/share their data.

Source: http://cgd.swissre.com/risk_dialogue_magazine/Healthcare_revolution/Data_Analytics_in_life_insurance.html

Aviva in USA

- Aviva USA had 60k life insurance applicants which it had underwritten in the traditional way – including blood and urine tests –and categorised accordingly.
- Deloitte took 30k applications and built a predictive model based on insurance application forms, industry information (past insurance applications and motor vehicle reports) and consumer-marketing data from Equifax Inc (hundreds to attributes per individual e.g. hobbies, income, TV-viewing habits).
- Tested predictive model on other 30k to see if could replicate underwriters' traditional assessments.
- "The use of third-party data was persuasive across the board in all cases," said John Currier, chief actuary for Aviva USA

Source: http://www.wsj.com/articles/SB10001424052748704104104575622531084755588

Non-Actuaries Outperforming Actuaries in this Field

HCF customer retention initiative

- In 2013, Australian health insurer HCF (through Deloitte) invited data scientists to analyze their data to identify policyholders most likely to lapse
- 300 data scientists from Kaggle were invited from around the world from which three submissions were selected for closer examination to use in building a "predictive algorithm that allows them to tailor their health cover more closely to member needs"

Liberty Mutual fire loss prediction

- In 2014, Liberty Mutual ran a contest on Kaggle to predict fire losses to enable more accurate assessment of policyholder's risk exposure
- 634 entries were submitted included 19 from Liberty Mutual employees. The best Liberty Mutual entry was ranked 36th in the competition
- In a similar competition run by Allstate in 2011, the participants were able to achieve a 340% improvement over Allstate's ability to predict bodily injury insurance. And that too, with anonymized data and not knowing true makes and models of the cars.[1]

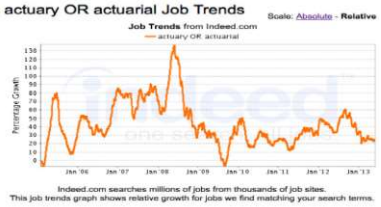
Past member of our working party, Xavier, won both of these competitions demonstrating that it is possible for actuaries to excel in this field

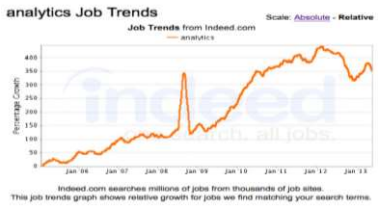
[1] **Source**: http://andrewmcafee.org/2012/03/a-data-scientist-youve-never-heard-of-is-now-the-master-of-your-domain/

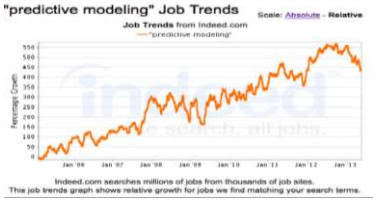
Job Trends & Employer Demand

- Demand for traditional actuarial roles expected to remain strong in Asia driven by market growth and regulatory developments
- Outside of Asia, in developed markets, predictive modelling and analytics are growing much faster than traditional actuarial jobs. This trend may extend to Asia in the long term.

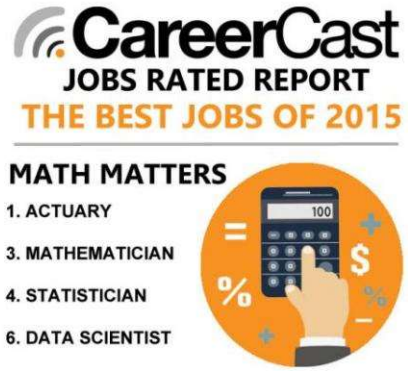
Demand for actuarial jobs flat while growth in analytics and predictive modelling jobs







Source: Presentation by Morand & Troceen, DW Simpson at ICA 20141

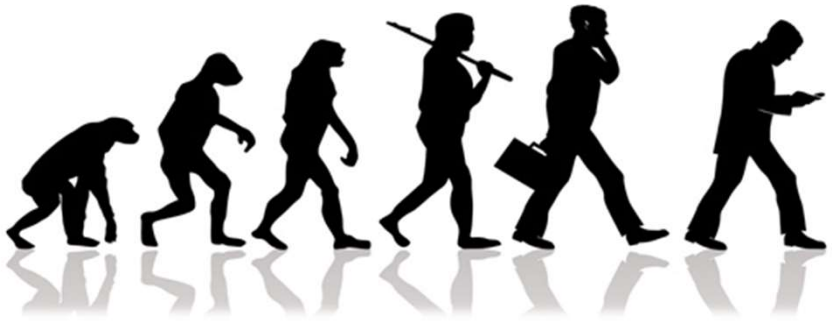


Four math related jobs in Top 6

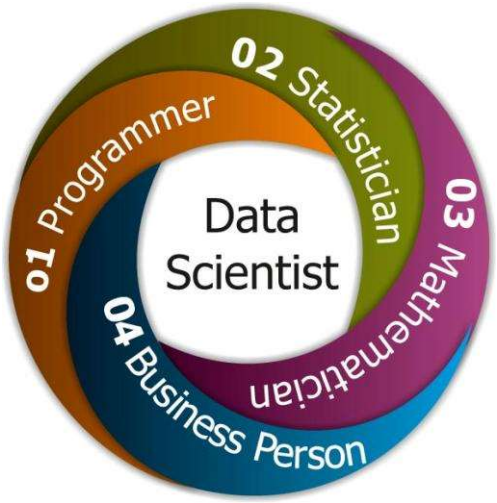
http://www.careercast.com/jobs-rated/best-jobs-2015

Actuaries of the Fifth Kind?

Hans Bühlmann 1987	Actuaries of the First Kind	• 17th century: Life insurance, Deterministic methods
	Actuaries of the Second Kind	• Early 20th century: General insurance, Probabilistic methods
	Actuaries of the Third Kind	• 1980s: Assets/derivatives, Contingencies Stochastic processes
Paul Embrechts 2005	Actuaries of the Fourth Kind	• Early 21st century: ERM
Big Data Working Party	Actuaries of the Fifth Kind	• Second decade of 21st century: Big Data



Skills Required



Source: http://www.edureka.co/blog/who-is-a-data-scientist/

Actuaries

- Possess good computing skills
- Are good at math & statistics
- Have deep understanding of business

Actuaries as managers or modelers have a niche in the data science arena

Need to upgrade skillset with emerging tools and techniques relevant to analyze big data

- **Management**: to understand the process, what questions to ask, what skillset to hire
- **Modelling**: to build skillsets that are growing in importance

Need to Learn New Tools

Harvard Business Review advise to managers hiring data scientists[1]:

"Don't bother with any candidate who can't code"

Excel is excellent for learning & visualization but has limitations

- Data size
- Complex analysis becomes difficult (e.g. GLMs)

Tools for big data analytics

- Good first step: R, Python
- Longer term: Revolution R, Hadoop, Microsoft Azure, DataRobot

Useful references

- For a detailed comparison of software options, see presentation by Hugh Miller: http://www.actuaries.asn.au/Library/Events/GIS/2014/5CMillerSoftwarePres.pdf

[1] **Source**: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

Where to Begin?

Beginner resources

- Lots of online resources
- Attend the workshop

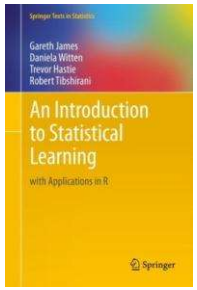
Online courses

- Online course by Caltech: https://work.caltech.edu/telecourse.html
- Online course by Andrew Ng, Stanford University: https://www.coursera.org/course/ml

iTunes U coursera

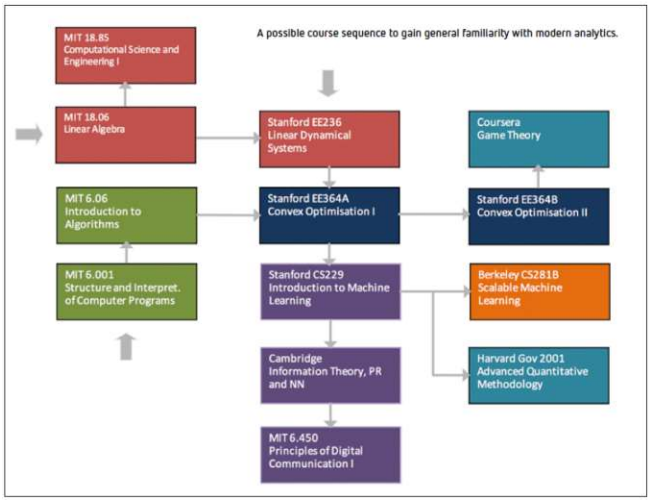
Textbook

- An Introduction to Statistical Learning with applications in R: http://www-bcf.usc.edu/~gareth/ISL/



In depth learning

- Dimitri Semenovish provides a sample learning pathway (shown below) using courses available online
- Refer to his article in Actuary Australia for more detail: http://actuaries.asn.au/Library/AAArticles/2014/Actuaries191JULY2014p22t25.pdf



Data analytics case study

Part II

Asia Actuarial Analytics Challenge 2016

Predict re-admission to hospital for diabetes patients

58 teams, 70 players and 578 entries

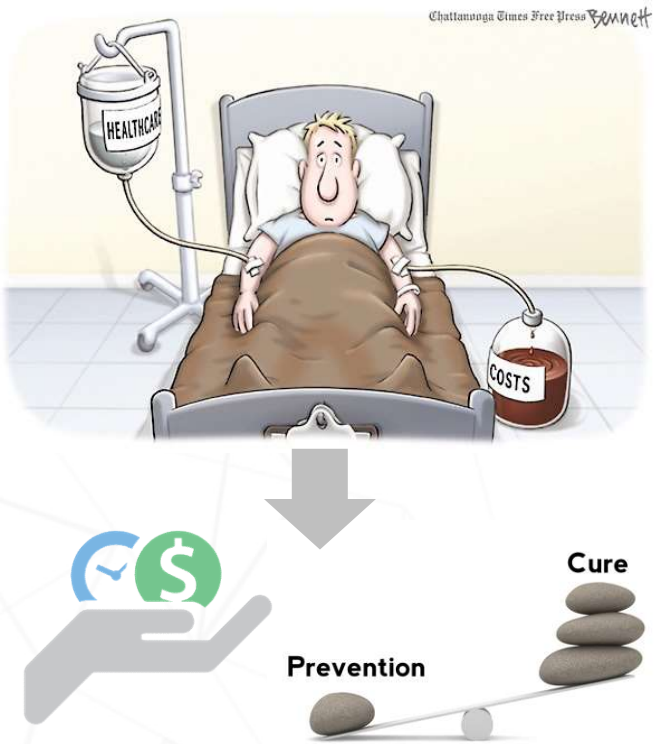
Rank	Team	Score	Country
1	Holmusk	0.69272	
2	Saliya J	0.68139	
3	lh_teh	0.67890	


holmusk

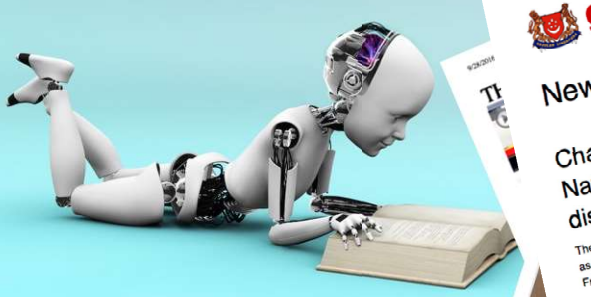

Machine Learning

Why is this challenge important?

Identify high-risk readmission patients



Machine learning
is the future of analytics



Bachelor of Computing

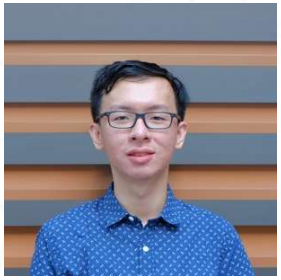
Medical Doctor

Master in Bioinformatics

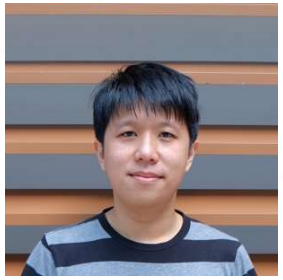
Master in Statistics

Ph.D. in Electrical Engineering

Ph.D. in Electrical Engineering



Bachelor of Physics



Bachelor of Actuarial Science

Master in Machine Learning

Brief intro of our team:

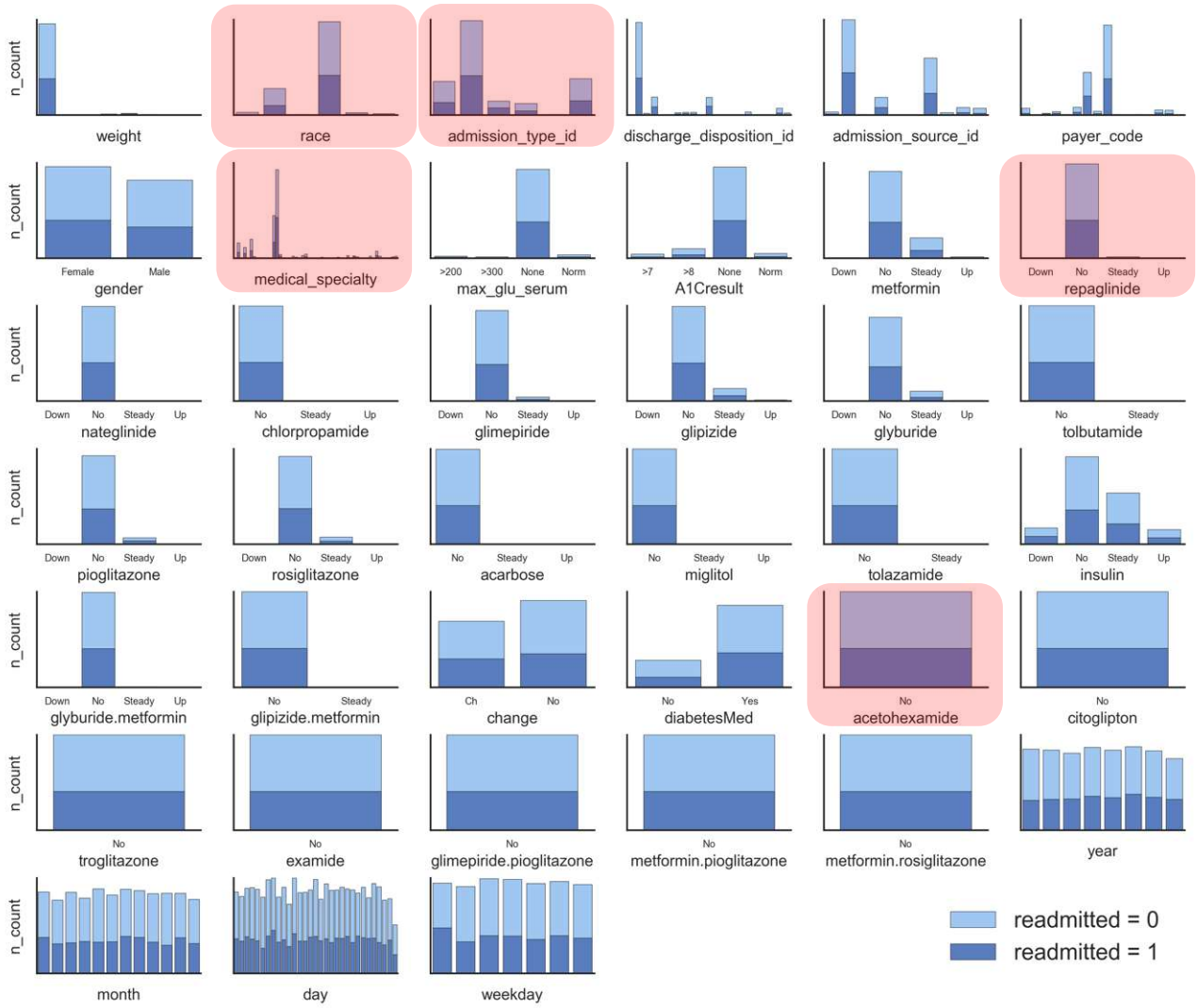
We are from diverse background such as actuarial, medical, bioinformatics, physics, computer science, statistics and machine learning

With in-house doctor's medical opinion, machine learning pipelines and data engineering modules, these areas of expertise have been very helpful in this challenge

Overview of what is in the data

- Categorical variables (e.g. race)
- Numerical variables (e.g. age)
- Diagnosis information (e.g. ICD9 code)

Categorical variables



Stacked bar chart for categorical variables

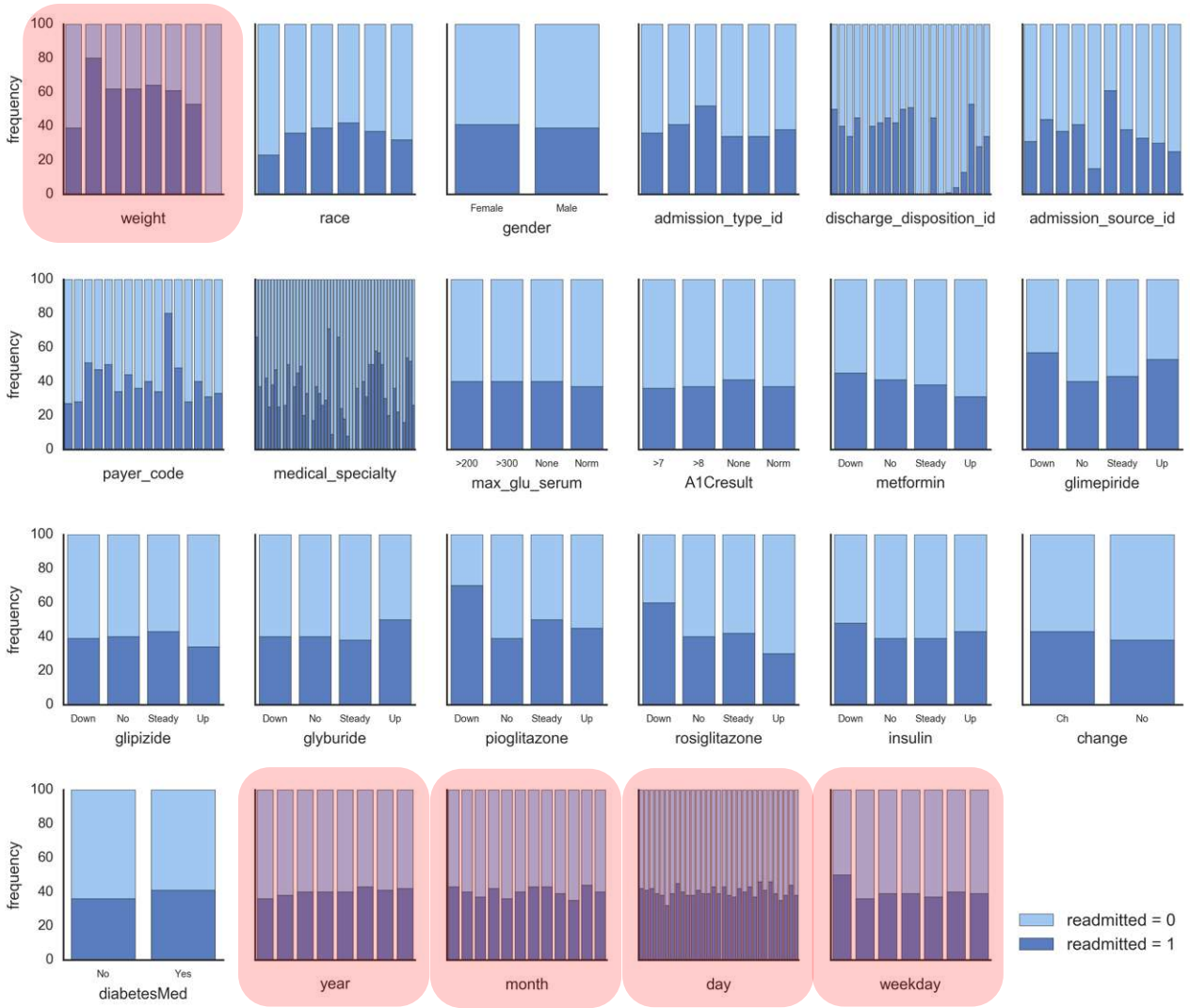
Race
- Caucasian
- African American

Admission_type_id
- Emergency
- Urgent
- Elective

Medical Specialty
- Cardiology
- Radiologist

Observation

Some variables are monotonous or dominated by ONE categorical value. These variables are quite redundant (Appendix 1).

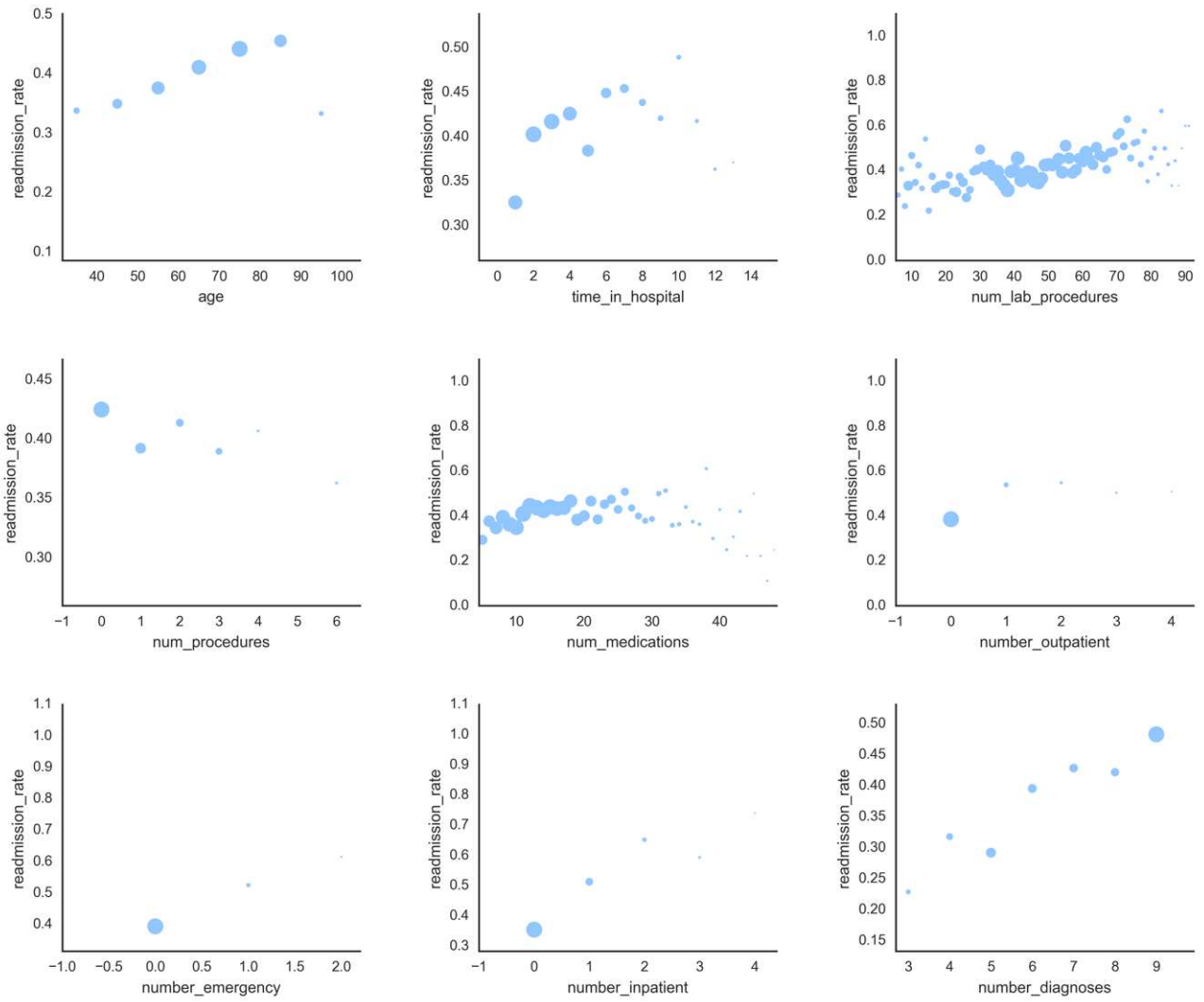


100% stacked bar chart to reflect readmission rate

Observation

- Weight variable seems to have high predictive power on readmission rate
- [Day], [Month], [Year] & [Weekday] information extracted from patient's admission date DD/MM/YYYY seems to have some trend and correlation with readmission rate (we will look into it later)

Numerical variables

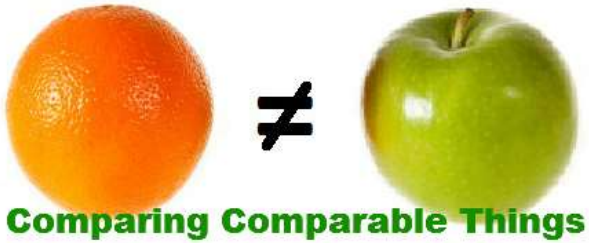


What we can observe from numerical variables:

- Some correlates with readmission rate
- Some have non-linear relationship with readmission rate
- Hence, we should also consider using **non-linear machine learning models** such as,
  - Decision tree
  - Random Forest
  - Gradient Boosting

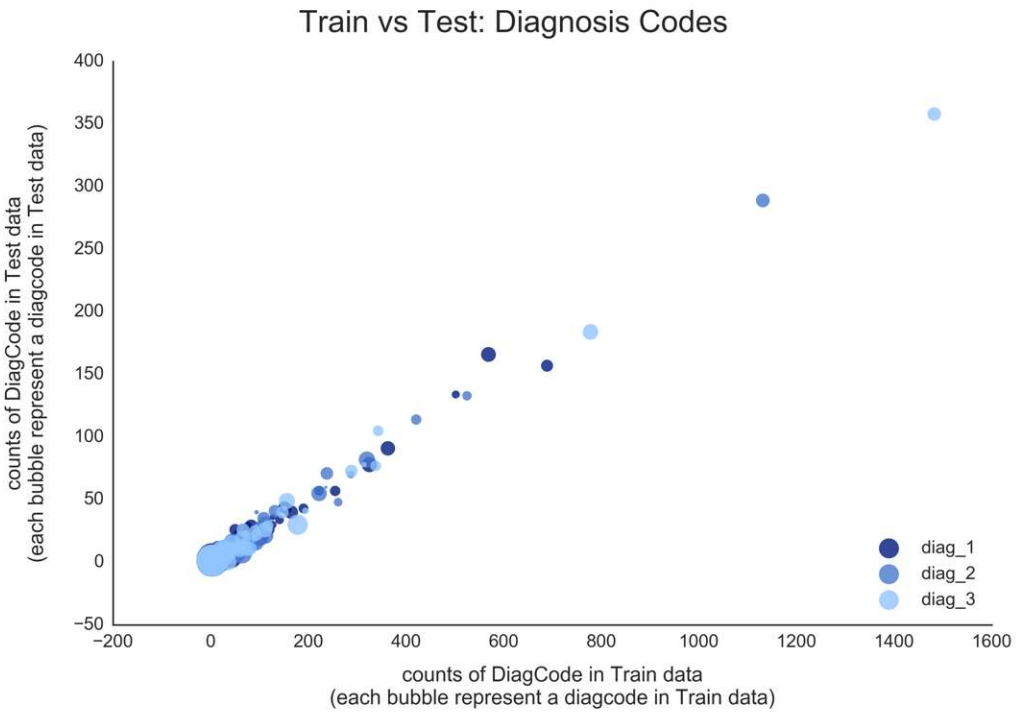
Diagnosis Information

We can use ICD9 code & diagnosis description to do preliminary check if patients are similar in both training and testing data set



If they are not, we will need to handle the problem a little bit differently by creating a validation set[2] which is a good representative of testing set

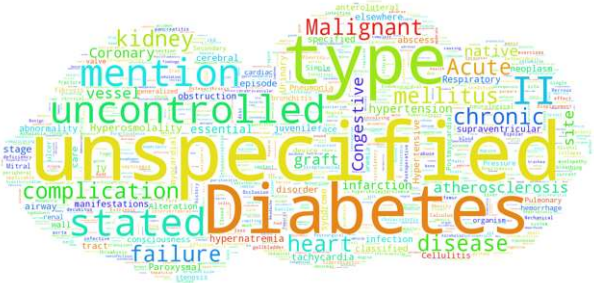
[2] Suggested by Colin in forum, code: https://github.com/zygmuntz/adversarial-validation



Patients from training & testing data set have similar medical conditions

Diagnosis Descriptions

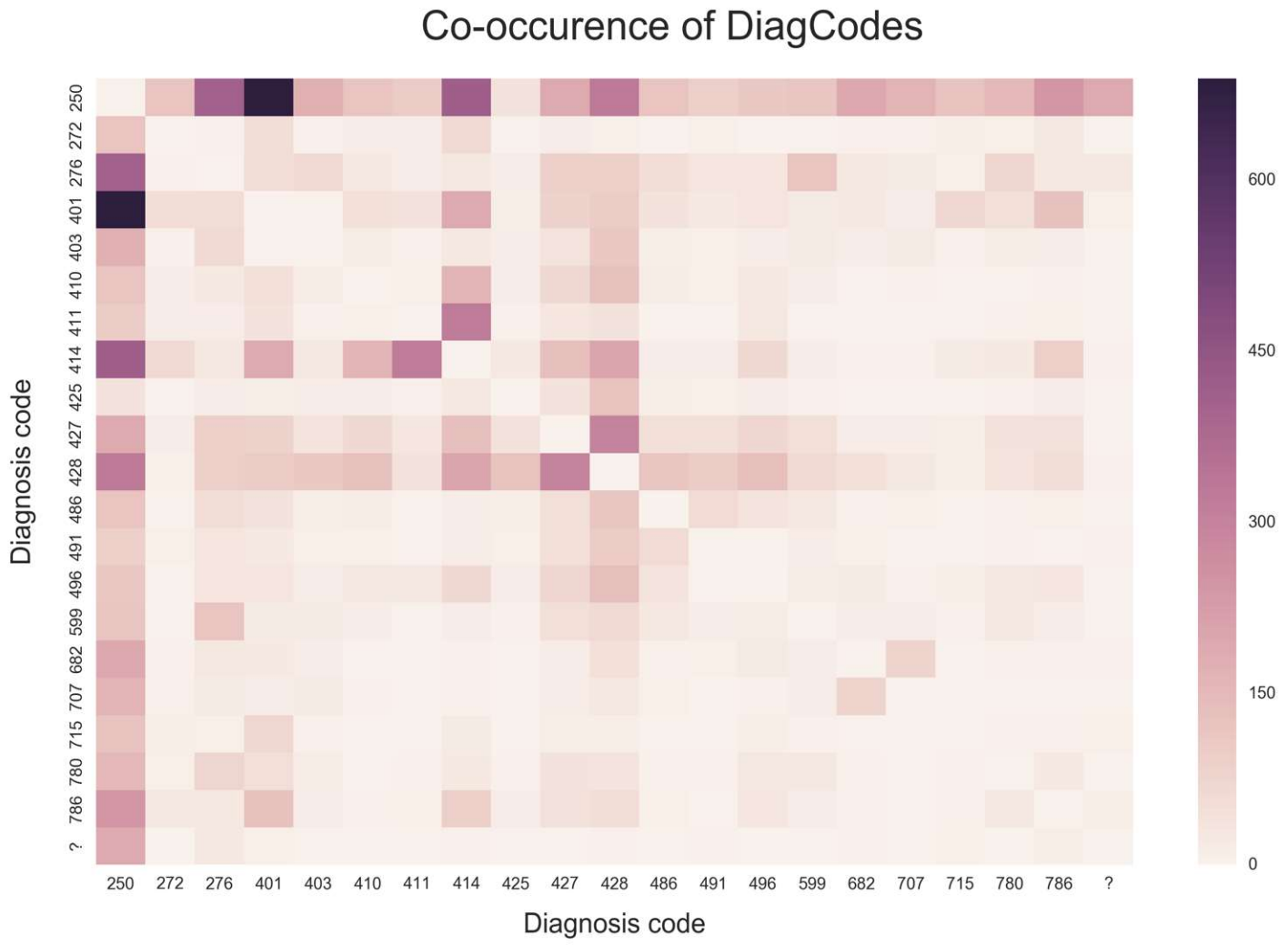
Training set



Testing set

They have similar key words in diagnosis description as well!

In this case, training data is a **good representative data set** for testing data



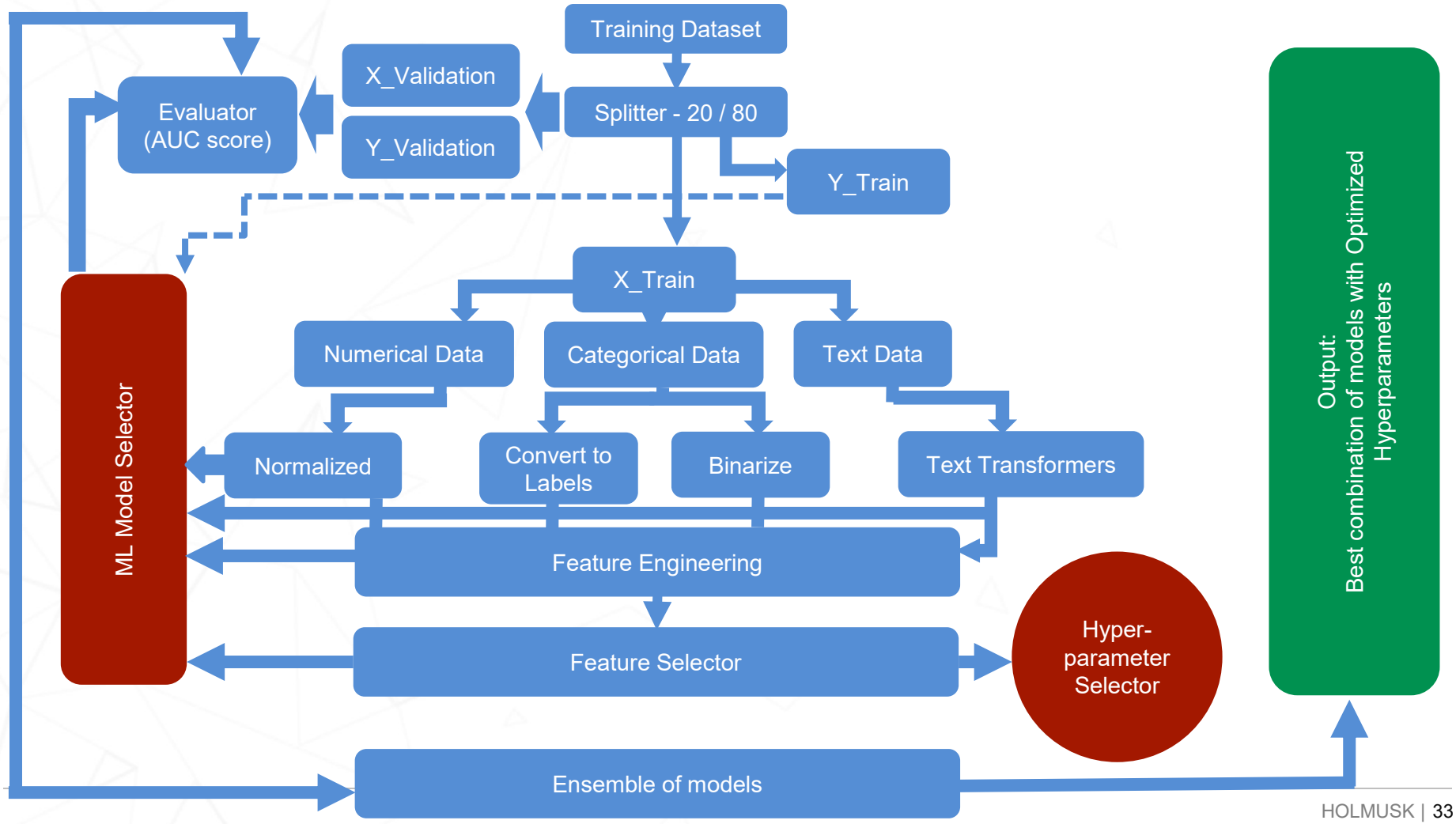
From diagnosis code information:

- We can also do comorbidity study, here is the heatmap that we can see how often certain diagnoses appear with each other
- We can use these information to reduce the dimension of data set, i.e. by grouping related medical conditions

Machine Learning Modelling

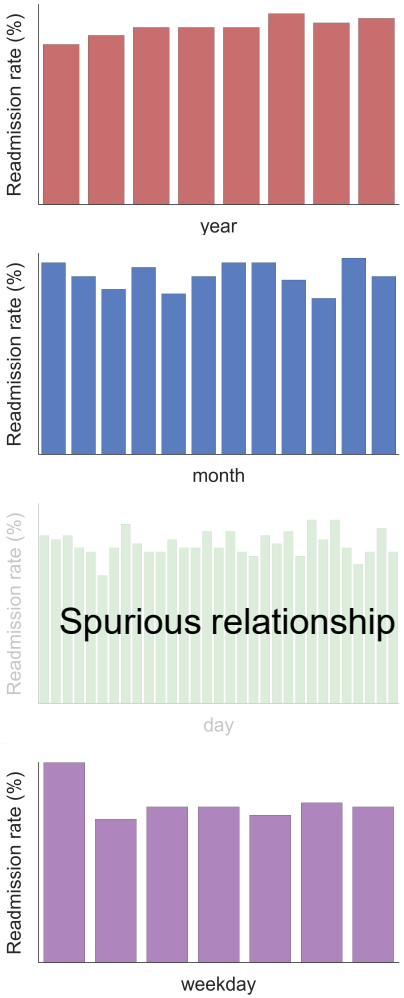
- Process flow diagram
- Feature Engineering & Study
  i. Admission Date
  ii. Diagnosis Codes
  iii. Diagnosis Descriptions
- Model Selection

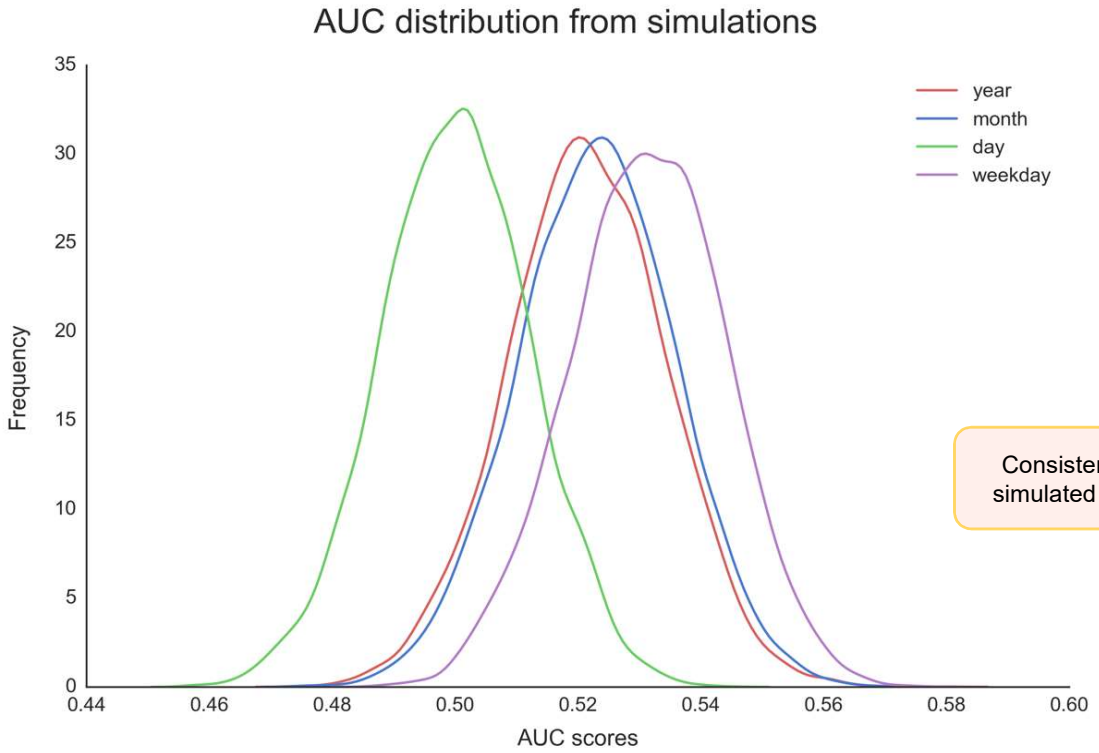
Process flow diagram


Training Dataset
X_Validation
Evaluator
(AUC score)
Splitter - 20 / 80
Y_Validation
Y_Train
X_Train
Numerical Data
Categorical Data
Text Data
ML Model Selector
Normalized
Convert to
Labels
Binarize
Text Transformers
Feature Engineering
Hyper-
parameter
Selector
Feature Selector
Output:
Best combination of models with Optimized
Hyperparameters
Ensemble of models

I. Feature Engineering: Admission Date

Handle spurious relationship for features **Year, Month, Day & Weekday** information extracted from admission date





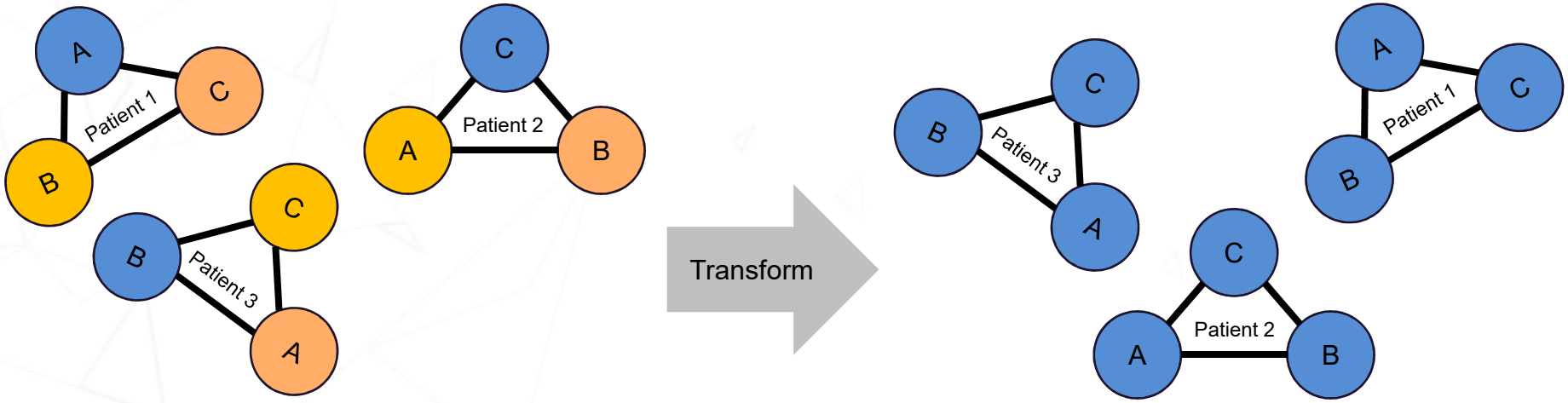
Consistent with simulated results

From Chi-square test:

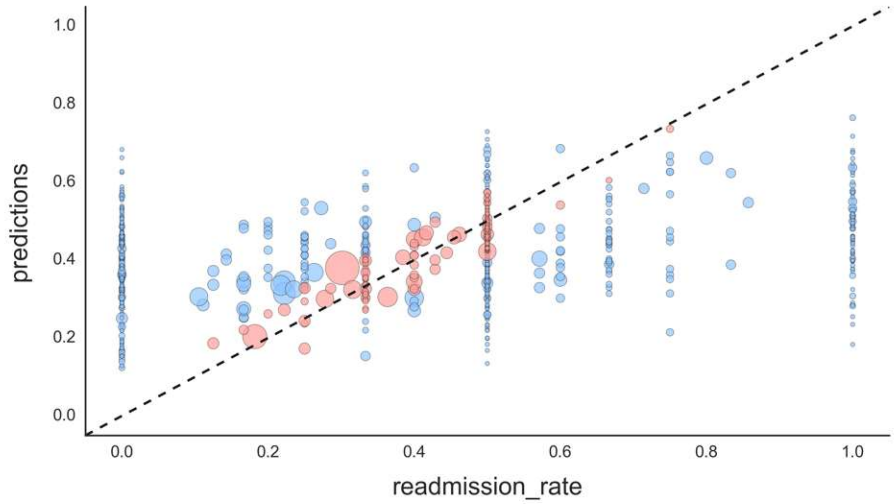
- Probability of 36% that the trend in **Year** happened by chance.
- Probability of 12% that the correlation in **Month** happened by chance.
- Probability of 96% that the correlation in **Day** happened by chance.
- Probability of ~0% that the correlation in **Weekday** happened by chance.

II. Feature Engineering: Diagnosis Codes

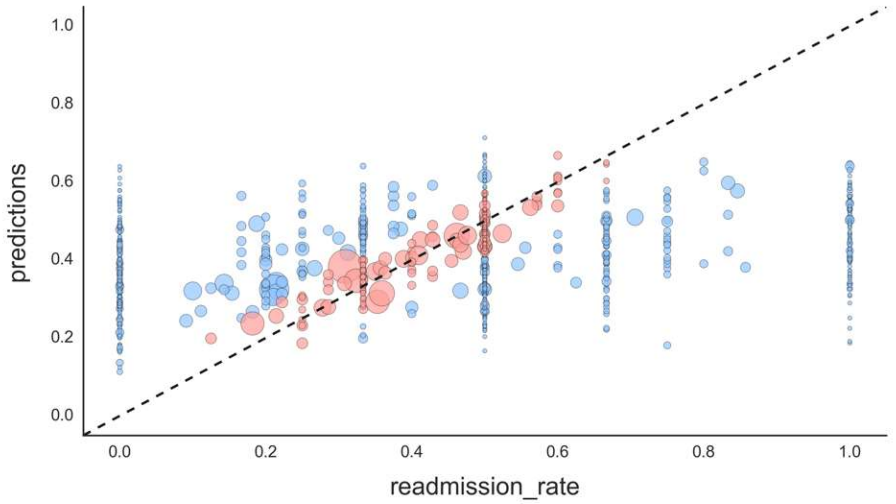
Transform diagnosis code sets with **order** to diagnosis code sets **without order**



Readmission_rate vs Preds - with order



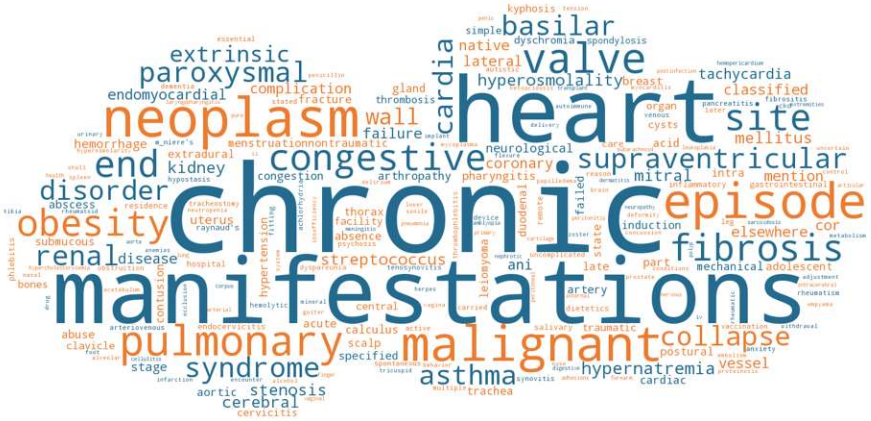
Readmission_rate vs Preds - without order



III. Feature Engineering: Diagnosis Desc

Diagnosis Description text mining

Word Cloud of important text found in diagnosis description



- Positive correlation with readmission rate
- Negative correlation with readmission rate

Example (patient #31758315):

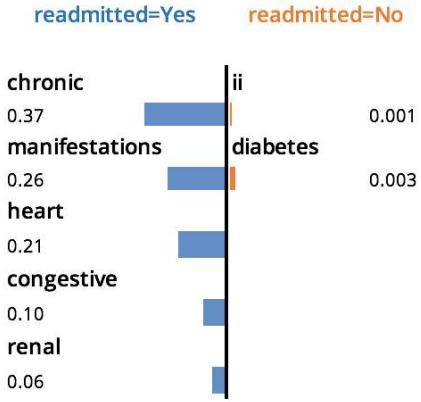
Text with highlighted words

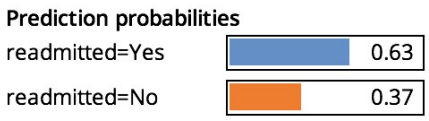
PatientId
31758315

Diag_1_desc
Congestive heart failure, unspecified

Diag_2_desc
Chronic kidney disease (ckd)

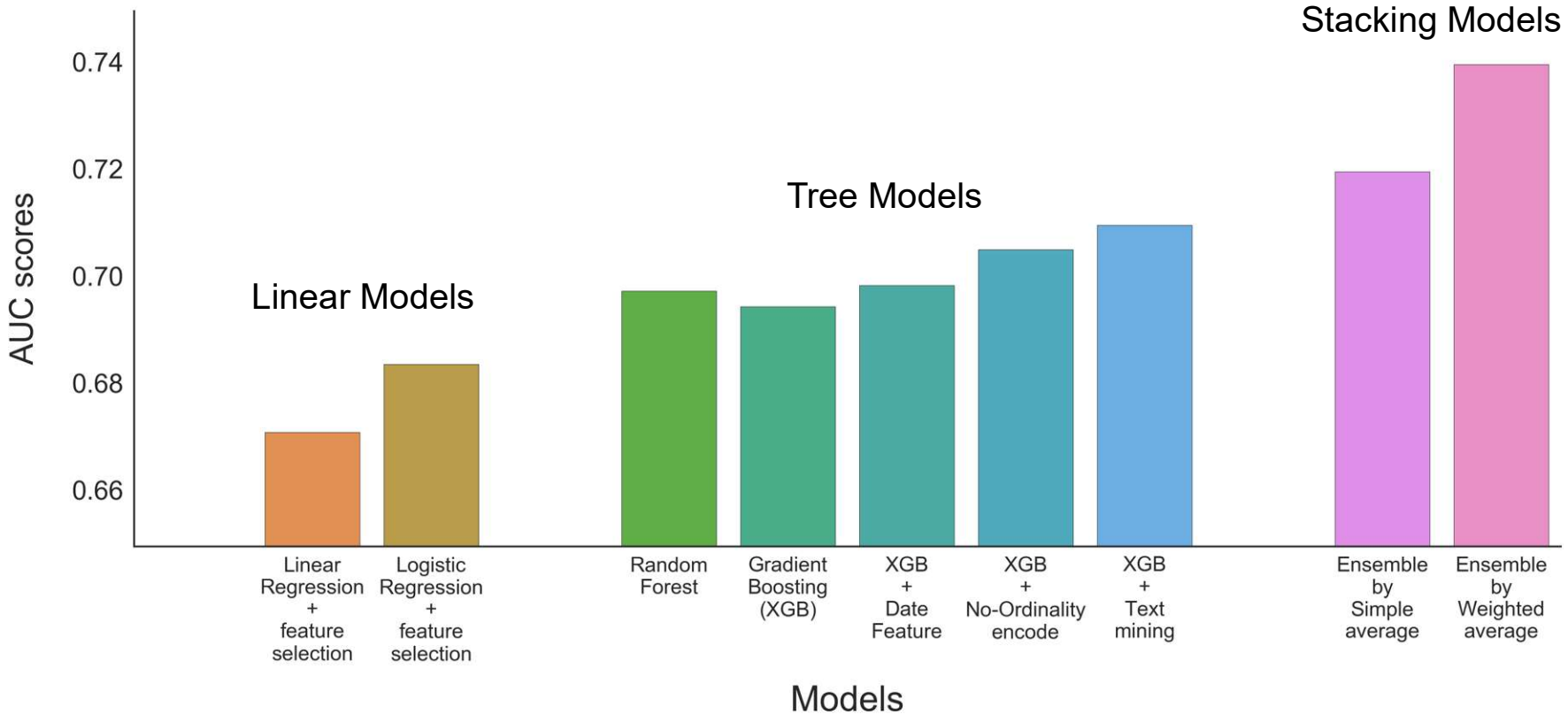
Diag_3_desc
Diabetes with renal manifestations, type II or unspecified type, not stated as uncontrolled





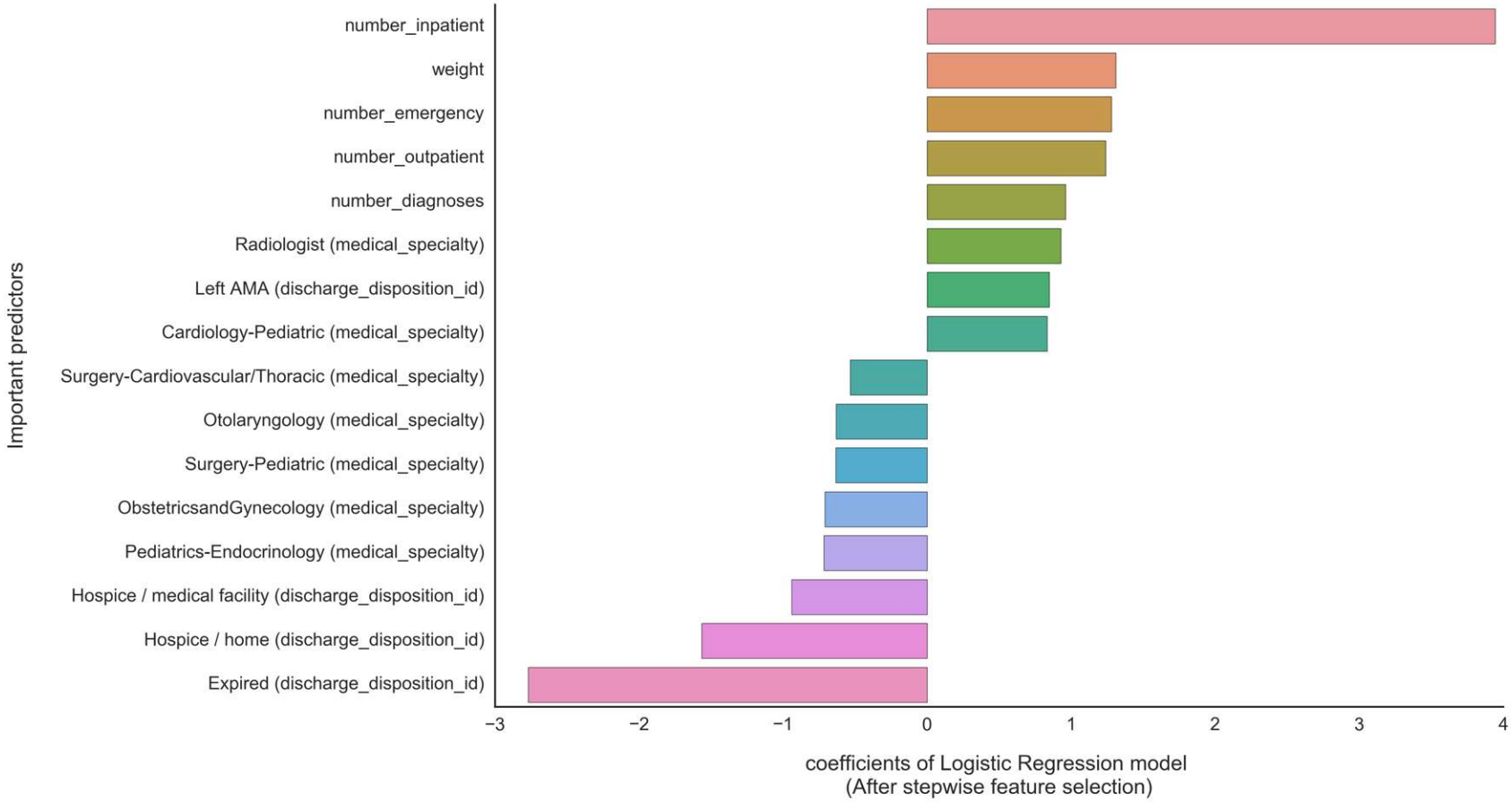
Model Selection

How does each model perform?



Findings

Feature Importance – Logistic Regression



Some key takeaways:

1. Feature engineering is the key!
2. When models work together, it gives even better result.
   (Stacking method)
3. Machine Learning can tell us something that we might not have expected.
4. Chronic disease such as diabetes involves many data from different sources beyond the ones in this challenge. By including these data, the model can be more robust in terms of disease early detection and prevention.
5. A mixed of domain expertise in a team can help a lot.

Q & A

Appendix

Appendix 1 - Preprocessing on categorical variables:

- Some variables are monotonous or dominated by ONE categorical value. Such as acetohexamide, repaglinide & etc.

  Hence, we removed these redundant variables

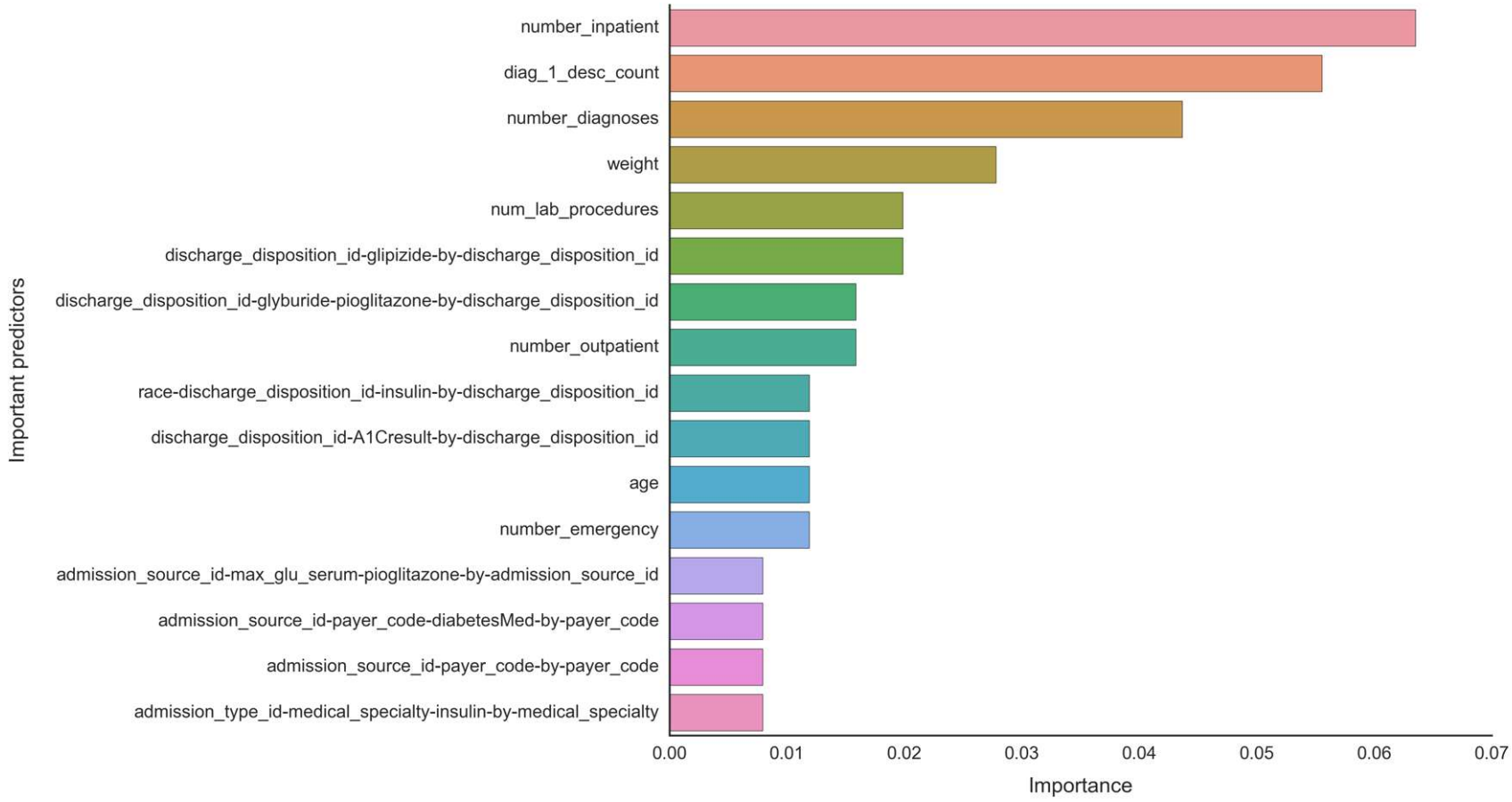
- For weight, 96% of the data is missing value but we didn't remove the weight. We will explain the reason why in next slide

Appendix 1 - Preprocessing on categorical variables

(100% stacked bar):

- Surprisingly, the 4% of weight data has a lot of information about readmission rate. Hence, weight variable is important to be included

Appendix 2 - Feature Importance, Gradient Boosting (Xgboost)



big data

technology

healthcare

holmusk

www.holmusk.com