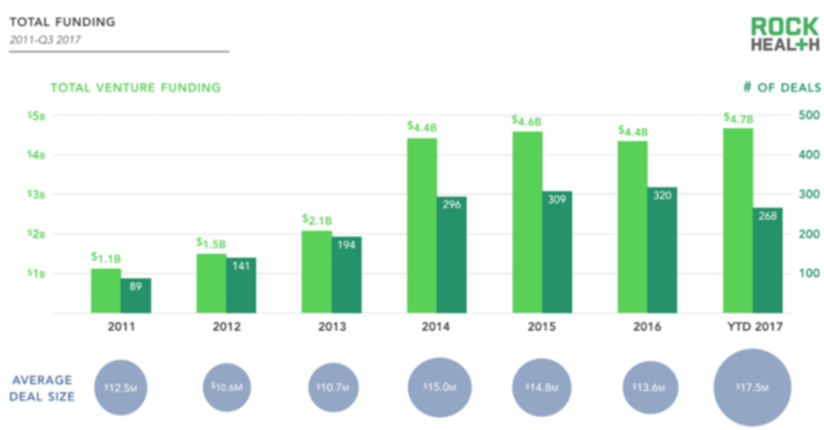
How DS is applied in healthcare industry?

Ella

Healthcare industry

- Explosive Spending in Health Care Tech
  - Health care market is a $3Trillion dollar industry



https://rockhealth.com/reports/2017-midyear-funding-review-a-record-breaking-first-half/

Examples of healthcare tech companies

SEVEN $100M+ VENTURE ROUNDS
H1 2017

ROCK HEALTH

Outcome HEALTH	*Consumer health information* Sells to: Providers	$500M Goldman Sachs, CapitalG	Chicago, IL
PELOTON	*Connected fitness equipment* Sells to: Consumers	$325M Wellington, KPCB, True Ventures	New York, NY
MODERNIZING MEDICINE	*EMR* Sells to: Physician practices	$231M Warburg Pincus	Boca Raton, FL
PatientPoint	*Consumer health information* Sells to: Providers	$140M Searchlight Capital Partners, Silver Point Capital	Cincinnati, OH
Alignment Healthcare	*Population health management* Sells to: Consumers, providers	$115M Warburg Pincus	Irvine, CA
patientslikeme	*Patient community* Sells to: Pharma	$100M iCarbonX	Cambridge, MA
sharecare	*Consumer health information* Sells to: Employers, health plans, providers	$100M Summit Partners	Atlanta, GA

Why data scientists are important in healthcare?

- Huge amount of data
  - 30% of the entire world's stored data is generated in the healthcare industry
  - A single patient typically generates close to 80 megabytes each year in imaging and electronic medical record (EMR).
- Healthcare is in strong need of data scientists
  - Of 6,000 data scientists in the US, only 180 are estimated to work in health care field (As of mid 2017)
  - There are nearly 6,000 hospitals and 400 academic medical centers, available labor force is a bit too thin

https://www.buildingbetterhealthcare.co.uk/technical/article_page/Comment_Health_networks__delivering_the_future_of_healthcare/94931

https://catalyst.nejm.org/case-data-scientists-inside-health-care/

How DS could help?

- Use cases in healthcare
  - Diagnostics
    - Detecting serious disorders or diseases using multiple data sources.
    - Improve hospital quality and patient safety
  - Prevention
    - Reducing preventable hospital readmission
    - Population health management, risk stratification, and prevention
  - Cash Flow Forecasting
    - Forecasting of cash flows based on claims history, reimbursement analysis and potential denials

https://healthitanalytics.com/news/four-use-cases-for-healthcare-predictive-analytics-big-data

How DS could help?

- Use cases in healthcare
  - Workflow Optimization
    - Using historical data for staffing to reduce costs, Having the right clinician at right time at right place
  - Efficient Use of Hospital Resources
    - Prevent bottlenecks in urgent care by analyzing patient flow during peak times
  - Grant problem
    - Predict likelihood that a particular proposal will receive grant using text analytics

What data is available?

- Codified Data Sets
  - Lab measurements
  - Bedside measurements (vital signs, ...)
  - Prescription orders, pharmacy fulfillment
  - Procedure and billing codes
  - Monitoring data
  - Intensive care
  - Home health
  - Genetics: SNPs, CNVs, Exomes, whole genome sequences
  - Geographic location

https://www.siam.org/meetings/sdm13/szolovits.pdf

What data is available?

- Narrative Data
  - Doctors' and nurses' notes
  - Radiology, pathology, ... reports
  - Discharge summaries
  - Referral letters
  - Blogs, diaries, posts to social media
- Imaging
  - MRI, scan and etc

Major steps to build a model

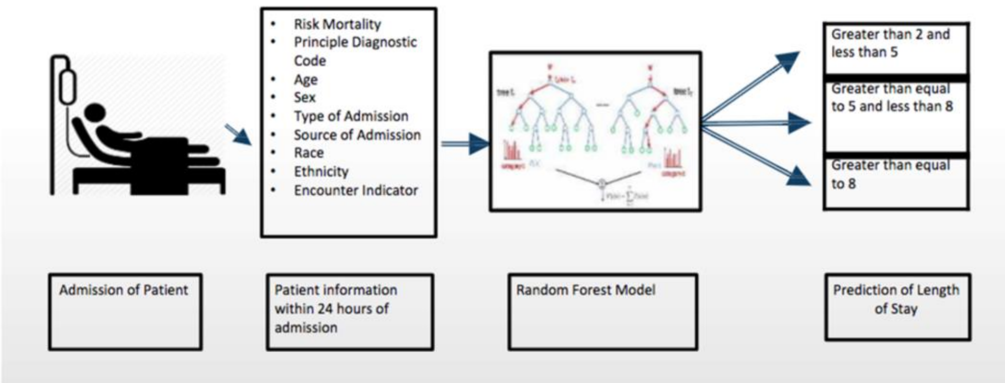
- Generate a large variety of features
  - Billing codes
  - Measured lab values
  - Medications and dosages
  - Frequency of doctors' visits and hospitalizations
  - Total "fact load"
  - NLP on notes and discharge summaries to find other evidence of the above
    - Results and prescriptions elsewhere are not in codified data, but are often mentioned in narrative reports

Major steps to build a model

- Data processing
  - Transformed variables: inverse, abs, square, square root, log-abs, abs deviation from mean, log, ...
  - Missing values: Some values are not measured for some clinical situations, Failures in data capture process, Episodically measured variables, Extrapolate, Unclear/undefined clinical state, Imprecise timing of meds.
- Feature engineering and selection
  - Derived variables can summarize essential contributions of dynamic variation: integrals, slopes, ranges, frequencies, etc.
- Machine learning algorithm

Problem 1

- Predict length of stay



Predicting length of stay (LOS)

- LOS
  - Defined as number of days from the initial admit date to the date that the patient is discharged from hospital.
- Source of variation
  - Patient condition
  - Various facilities
  - Specialties who treat patient
- Why it is important
  - enhance the quality of care
  - Improve operational workload efficiency
  - accurate planning for discharges resulting in lowering readmission

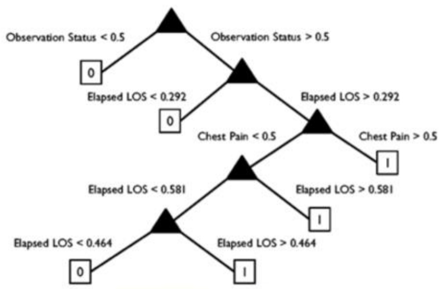
Predicting LOS workflow

- Features
  - The demographic and clinical predictors are static model inputs that are known at the time of admission.
  - Other predictors such as patient census, day of the week, and elapsed length of stay are dynamic and are continuously updated during a patient's stay.
- Response
  - Whether the patient was discharged by 2 p.m or by end of day
- Build models
  - Logistic Regression
  - Tree based models

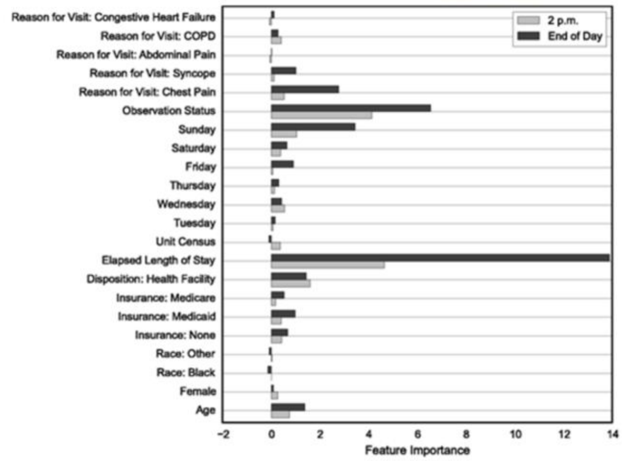
https://academic.oup.com/jamia/article/23/e1/e2/2379761

https://www.datasciencecentral.com/profiles/blogs/5-machine-learning-research-studies-to-understand-predict-length

Model fitting and output



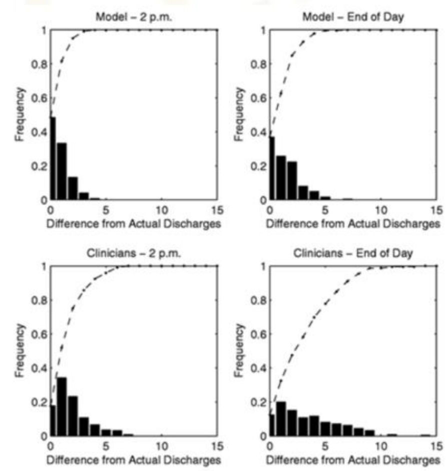
One example tree



Feature importance

Model performance

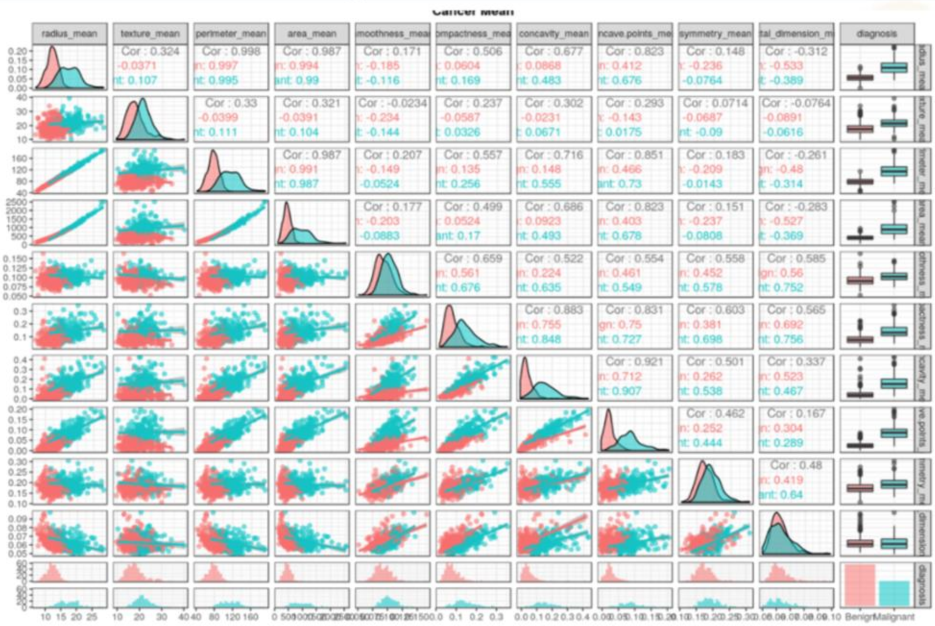
- Regular metrics
  - ROC, AUC, Precision Recall, F1 score
- Compare with benchmark model



Problem 2

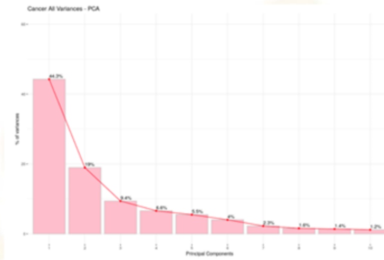
- Breast cancer prediction
  - Goal: Predict whether the cancer is benign or malignant
  - Data: Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.
- Features extracted from image
  - Radius, texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry, Fractal dimension
  - mean, standard error and "worst" or largest of above dimensions

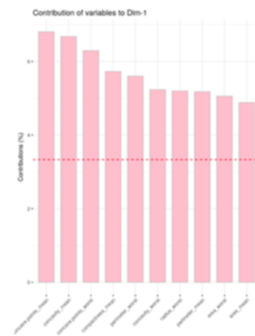
Correlation among variables



How to address collinearity

- Regularization
- Principal component analysis (PCA)
    - Choose appropriate number of components





    - How to name the new variables?
        - Plot contributions of top variables to each new dimension