2017

Internal Medicine

Keroles Hakem
Virginia Commonwealth University

Robert Trachy
Virginia Commonwealth University

Khanh Tran
Virginia Commonwealth University

COMPUTER SCIENCE

Internal Medicine

CAPSTONE DESIGN EXPO 2017

Project number: CS-318 | **Team members:** Keroles Hakem, Robert Trachy, Khanh Tran | **Faculty adviser:** Bartosz Krawczyk, PhD; Vimal Mishra, MD

Objective

Using machine learning, predict the length of stay of patients for MCV hospital based on the different character-istics of a patient, e.g., method of arrival. Design a web app for the end users based on the final predictive models.

Data Overview

Collected from MCV patients of the past 4 years
Contained over 130,000 patients and 66 features
Features include:

- Clinical characteristics: primary diagnosis, universal disease group
- Facility characteristics: bed category, admit unit, service provided, discharge disposition
- Socioeconomic factors: Admit source, insurance, method of payment

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

scikits learn

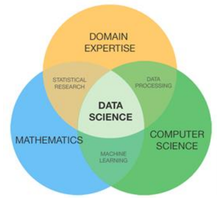
Approach

Predictive models are developed utilizing:

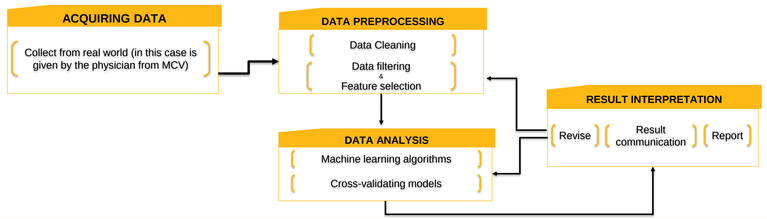
- Machine learning
- Statistical methods
- Professional input from domain experts.

Machine learning techniques using include:

- Supervised learning
- Regression models (linear, ridge, lasso)
- Cross-validation

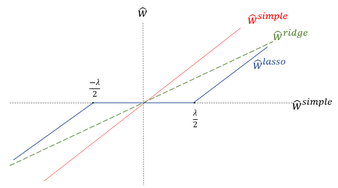


Method



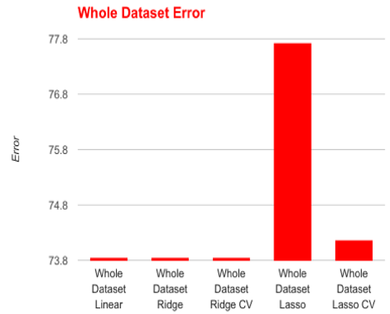
Linear vs. Ridge vs. Lasso Regression

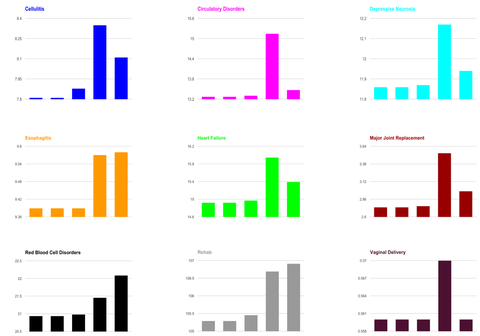
- **Linear Regression**: finding set of weights that would minimize the overall error between the predicted and actual outcome using gradient descent
- **Ridge Regression**: using L2 regularization to shrinkage weights, i.e., adds penalty equivalent to ***square of the magnitude*** of coefficients
- **Lasso Regression**: using L1 regularization that would exclude unimportant features from model, i.e., adds penalty equivalent to ***absolute value of the magnitude*** of coefficients.



Result & Prototype

The final data consisted of 9 features: *admit source*, *primary insurance*, *discharge disposition*, *admit unit*, *iso result*, *icu order*, *stepdown order*, *general care order*, and *age*. Ten different models for the top ten most frequent diagnoses and one model for the whole data set have been built. Data trained with and without *10-cross-validation*. Models are evaluated by the *mean square error* metric.





Prototype:

A web application is developed with Angular 2 framework:

- Friendly user interface that allows users to easily input information
- Form validation to ensure valid input
- Return an estimate of length of stays according to user input
- Report statistical summary of each model including weights, errors, and variances

Future:

In future research, more accurate prediction could be achieved with:

- Specific dataset of a group or similar diseases
- Larger dataset with more instances and potential factors
- Diverse dataset from multiple hospitals' records

Conclusion:

Our research project showcased the potential of using machine learning regression methods for length of stay prediction.

VCU School of Engineering