Stroke Risk – Project ISYE 7406

Joshua Mestemacher
Olan Malcolm
Phillip Sley
Syed Asghar

Spring 2025

AGENDA

- Problem Statement and Background
- Dataset and Exploratory Data Analysis
- Methodology
- Conclusion(s)

PROBLEM

Stroke is a leading cause of disability and death worldwide, making early prediction crucial for effective intervention and prevention.

Our objective is to leverage our data set to develop and evaluate machine learning models that can:

- Accurately predict an individual's likelihood of having a stroke.
- See what the most significant risk factors associated with having a stroke are.
- Tell us how age impacts an individual's stroke likelihood.
- Determine if there are any significant interactions between risk factors.

BACKGROUND

- In the United States, approximately 1 in 6 (17.5%) Cardiovascular related deaths are related to stroke[1].
- Every year approximately 795,000 people in the United States have a stroke[2].
- Stroke is the leading cause of long-term disability[2].
- Stroke-related costs in the United States came to nearly **$56.2 billion** between 2019 and 2020. Costs include the cost of health care services, medicines to treat stroke, and missed days of work[3].

[1] National Center for Health Statistics. Multiple Cause of Death 2018–2022 on CDC WONDER Database. Accessed May 3, 2024. https://wonder.cdc.gov/mcd.html
[2] Tsao CW, Aday AW, Almarzooq ZI, et al. Heart disease and stroke statistics—2023 update: a report from the American Heart Association. Circulation. 2023;147:e93–e621.
[3]Martin SS, Aday AW, Almarzooq ZI, et al.; American Heart Association Council on Epidemiology and Prevention Statistics Committee; Stroke Statistics Subcommittee. 2024 heart disease and stroke statistics: a report of US and global data from the American Heart Association. Circulation 2024;149:e347–913.

STROKE PREVENTION & INTERVENTION

- Modifiable Risk Factors of stroke include high blood pressure, improvement of Cholesterol levels, management or prevention of diabetes, and not smoking or having substance abuse[4].

- While there are modifiable risk factors, there are unmodifiable risk factors, including age, biological sex, family history, race and ethnicity, and prior stroke or heart attack[4].

- With modifiable risk factors, identifying a patient with increased stroke risk may help with prevention of stroke, leading to better public health outcomes.

[4] (2025) Prevention | National Institute of Neurological Disorders and stroke. Available at: https://www.ninds.nih.gov/health-information/stroke/prevention

PROBLEM STATEMENT (CONTINUED)

- Can we create a questionnaire for patient intake to appropriately triage patients for stroke?
- Can we create an accurate model for predicting stroke risk assuming precise medical measurements have been done?
- What are the most significant risk factors associated with a chance of having a stroke?
- How does age impact the likelihood of experiencing a stroke?
- Are there any significant interactions between risk factors predicting stroke?
- Traditional method (regression) versus ML (random forest, SVM) differences?

DATASET

- Our data set is the "Stroke Risk Prediction Dataset Based on Symptoms" from Kaggle.
- The data set consists of 18 columns:
  - There are 2 response variables – *Stroke Risk (%)* and *At Risk (Binary)*.
    - If someone was found to have a Stroke Risk ≥ 50% then they were classified as a 1 for At Risk.
  - The remaining 16 columns consist of variables that contribute toward an individual's risk of having a stroke.
    - 15 of the columns are symptoms that are binary (1 if the individual had it, 0 if not) and other column was the individual's age.
- Depending on what method was used, one of the response variables was chosen and the other was left out.

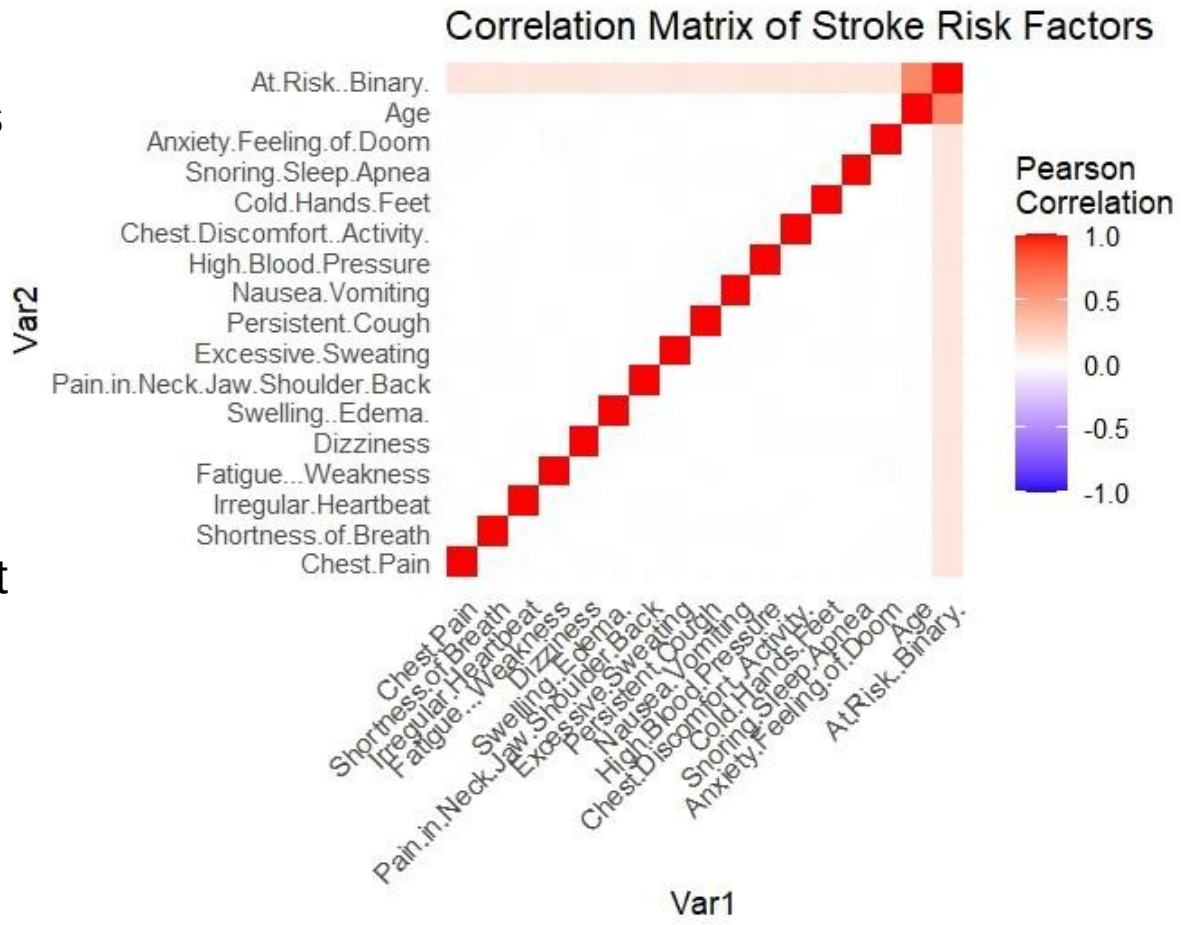
EXPLORATORY DATA ANALYSIS

- Chest Pain
- Shortness of Breath
- Irregular Heartbeat
- Fatigue Weakness
- Dizziness
- Swelling Edema
- Pain in Neck, Shoulder or Back
- Excessive Sweating
- Persistent Cough
- Nausea Vomiting
- High Blood Pressure
- Chest Discomfort
- Cold Hands or Feet
- Snoring or Sleep Apnea
- Anxiety or feeling of doom
- Age

- All independent variables in the data set are binary (0 or 1), except for Age (Continuous).
- There were no NULL values anywhere in the data.
- We explored the data to review linkage between each variable and the likelihood of stroke risk.
- We calculated a correlation matrix to view correlations between each of our variables within the model.
- We investigated possible multicollinearity using Variable Inflation Factor (VIF) to determine whether any variables should be removed due to multicollinearity.

EXPLORATORY DATA ANALYSIS

- Age seems to be a strong predictor for Stroke.
- The strong diagonal pattern with well-defined blocks suggests that variables are distinctly separable, which can lead to separation issues in our data set.
- The matrix shows relatively clear and distinct relationships without much noise. Real world medical data might show messier correlations.
- This "too clean" pattern can indicate that the dataset might have been generated using certain rules or criteria rather than capturing the natural variability found in real patient populations – Synthetic nature of dataset.



EXPLORATORY DATA ANALYSIS

- Age has a correlation of 0.612, much higher than all other predictors for stroke.
- The Variance Inflation Factor (VIF) based off a simple logistic regression shows that all predictors have very high values – reinforces Multicollinearity issue.
- Nearly identical VIF values (all around 100) is unusual and suggests they might be systematically related through some formula or calculation method – Synthetic nature of data issue.
- Furthermore, a split of 30% train and 70% test, instead of the reverse for the baseline logistic regression model also gave 100% accuracy results. This reinforced that Age was a perfect predictor of stroke risk.

➢ We shall remove **Age** as a predictor – it seems to predict stroke risk too perfectly.

➢ We will do **feature engineering** to explore which set of predictors can best be used for modelling to help our problem statement.

Correlation of factors with Stroke Risk

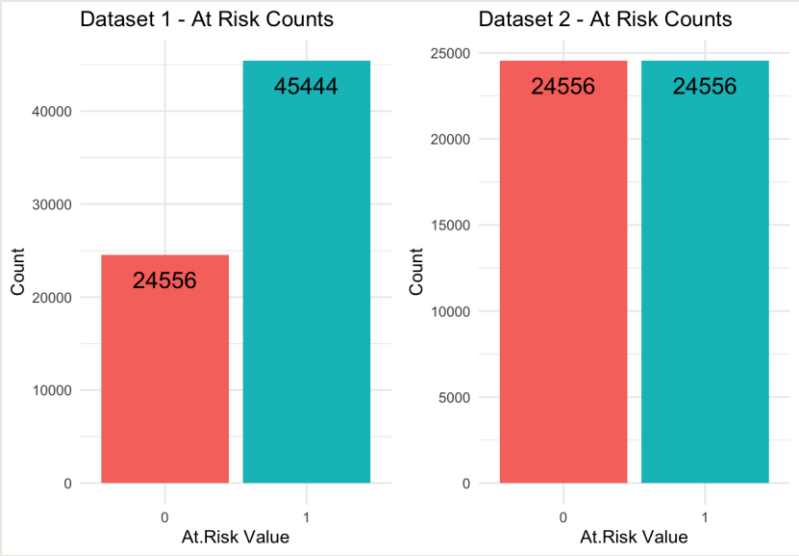
Age	Cold.Hands.Feet	Chest.Pain
0.6120384	0.1366420	0.1353654
Snoring.Sleep.Apnea	Fatigue.Weakness	Excessive.Sweating
0.1336813	0.1330595	0.1328057
High.Blood.Pressure	Anxiety.Feeling.of.Doom	Shortness.of.Breadth
0.1323011	0.1314662	0.1304822
Dizziness	Persistent.Cough	Nausea.Vomiting
0.1304435	0.1287005	0.1286452
Irregular.Heartbeat	Chest.Discomfort.Activity	Swelling.Edema
0.1241500	0.1236401	0.1224162
Pain.in.Neck.Jaw.Shoulder.Back		
0.1202803		

Variable Inflation Factor (VIF) (Log. Regression)

Chest.Pain	Shortness.of.Breadth	Irregular.Heartbeat
100.6360	100.1477	100.6785
Fatigue.Weakness	Dizziness	Swelling.Edema
100.6051	100.7278	100.7230
Pain.in.Neck.Jaw.Shoulder.Back	Excessive.Sweating	Persistent.Cough
101.0556	100.8956	100.4156
Nausea.Vomiting	High.Blood.Pressure	Chest.Discomfort.Activity
101.0275	100.5353	100.7581
Cold.Hands.Feet	Snoring.Sleep.Apnea	Anxiety.Feeling.of.Doom
100.6392	100.6465	100.7107
Age		
1004.2275		

MAIN METHODOLOGY - APPROACH

- The data was not balanced – there were almost double the amount of 1 At Risk responses (45,444) than there were 0's (24,556).
  - To balance the data set, we took a sample of the 1 responses to equal the amount of 0 responses in our dataset – both response variables now had an equal amount of 24,556 observations.
  - This was necessary so that there was not a bias towards At Risk/higher Stroke Risk percentage for the models ran.
- Split the data into train (70%) and test (30%) data sets.
- Fit some baseline models (logistic regression, ridge regression, naïve bayes, boosting, random forest, etc. with all remaining predictors) and review results like Accuracy, Precision, MSE, R-squared as appropriate.
- Perform feature selection based on results.
- Refit model with optimal selected set of features.
- Cross validate performance.
- Report results/findings.



MODELS USED

Traditional Statistical Methods

- Logistic Regression
- Firth’s Regression
- LASSO Regression
- Elastic Net Regression
- Ridge Regression
- Beta Regression

Machine Learning Methods

- Decision Tree
- Random Forest
- Gradient Boosting
- K-Nearest Neighbors (KNN)
- Naïve Bayes

LOGISTIC REGRESSION

A supervised learning algorithm for binary classification. It estimates the probability of an outcome using the logistic function and minimizes the log-loss function.

- **Why used?** A fundamental model for binary classification problems – applicable in our case where **At Risk (0 or 1)** was used as the response variable for out models.
- Simple and highly interpretable.
- **Relevance to Stroke Risk Dataset:**
- Establishes a direct relationship between symptoms and stroke risk.
- Provides interpretable coefficients to understand symptom importance.

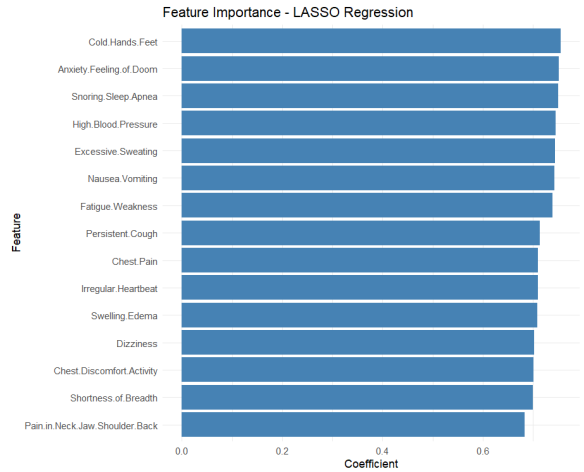
FIRTH'S LOGISTIC REGRESSION

A variation of logistic regression that applies a bias-reducing penalty to the likelihood function to prevent issues from complete separation.

- **Why used?** Addresses separation issues, especially when Age perfectly predicted the outcome, necessitating removal.
- Slightly more complex but interpretability still high.
- **Relevance to Stroke Risk Dataset:**
- Helps mitigate bias in a dataset where symptoms independently contribute.
- Improves prediction stability when standard logistic regression struggles.

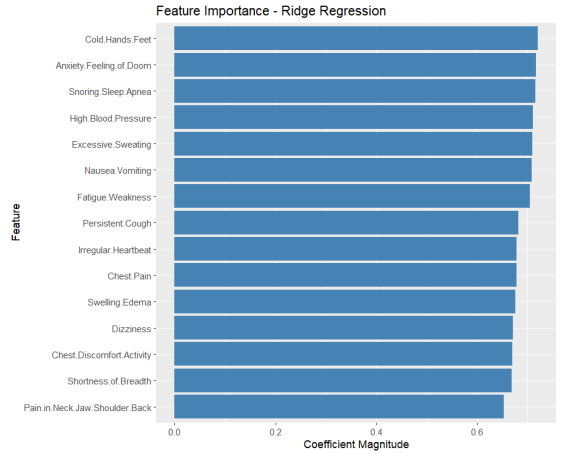
LASSO REGRESSION

- A supervised learning method that applies L1 regularization, which minimizes the sum of absolute coefficient values, effectively selecting important features.
- **Why used?** Performs feature selection via L1 regularization.
- Helps prevent overfitting and enhances model simplicity but may discard some useful predictors – which it did not in our case.
- **Relevance to Stroke Risk Dataset:**
- Helps identify the most critical symptoms for predicting stroke risk.
- Useful in reducing the dimensionality of the dataset.
- Reduces overfitting by setting insignificant feature coefficients to zero.



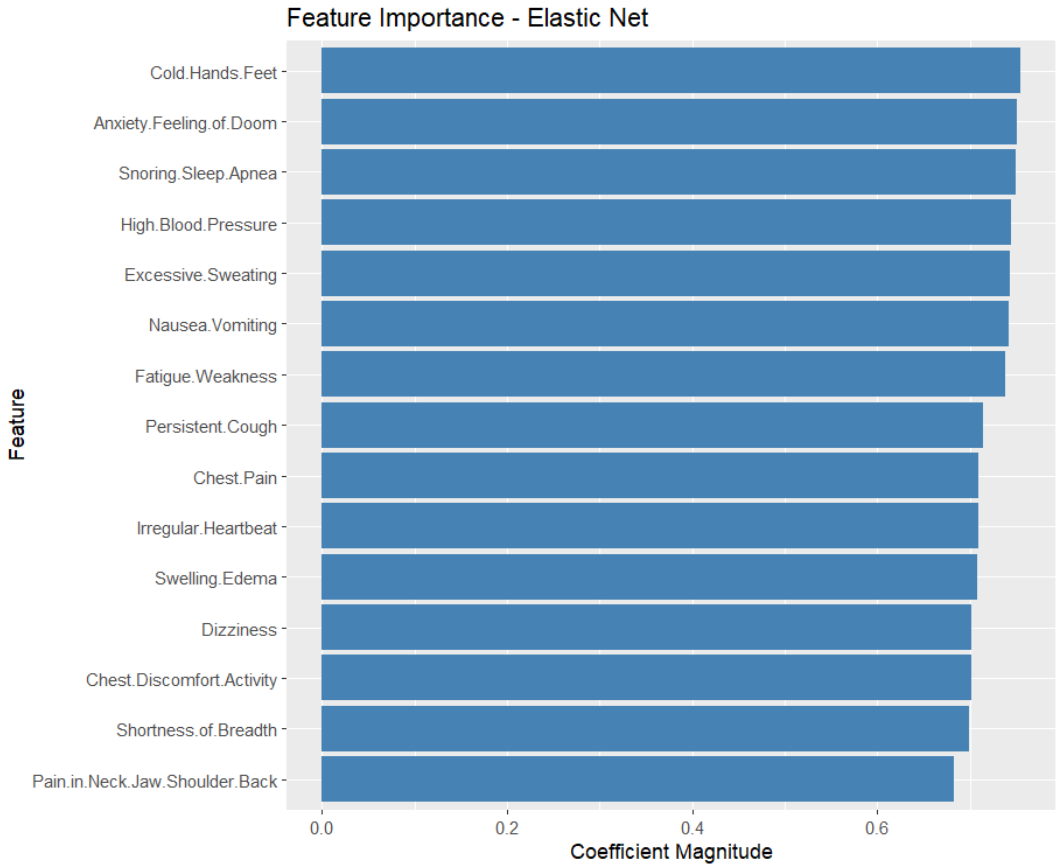
RIDGE REGRESSION

- A supervised learning algorithm that applies L2 regularization, minimizing the sum of squared coefficients to prevent overfitting while maintaining all features.
- **Why used?** Reduces overfitting using L2 regularization.
- Slightly more complex but interpretability still high.
- Generalizes well by preventing extreme coefficients but does not perform feature selection.
- **Relevance to Stroke Risk Dataset:**
- Stabilizes predictions when many features contribute to stroke risk.
- Avoids extreme coefficient values by distributing importance across features.
- Retains all symptoms as predictors, stabilizing their importance.



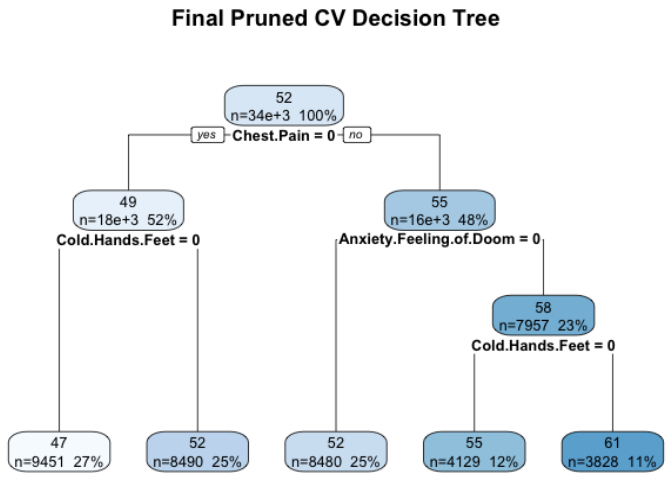
ELASTIC NET REGRESSION

- A supervised learning approach that combines L1 (Lasso) and L2 (Ridge) regularization to balance feature selection and coefficient shrinkage.
- **Why used?** Combines L1 (Lasso) and L2 (Ridge) penalties for balanced feature selection and regularization.
- Generalizes well in cases with correlated features and helps balance between overfitting and underfitting.
- Simple and highly interpretable.
- **Relevance to Stroke Risk Dataset:**
- Captures complex relationships between correlated features – given the issue of perfect separation and synthetic nature of dataset, seems applicable.
- Balances feature selection and predictive stability, making it ideal for correlated symptoms.
- Ideal for datasets where symptoms are expected to be interrelated.



DECISION TREE

- A supervised learning algorithm used for predicting continuous values – was used to predict Stroke Risk percentage.
- Works by splitting data into smaller subsets based on conditions that minimize variance.
- Consists of decision nodes (splits) and leaf nodes (predicted values).
- **Why used?**
  - Easy to interpret and visualize.
  - Handles nonlinear relationships well.
  - Can handle different types of data (numerical and categorical) without requiring extensive preprocessing.



- **Relevance to Stroke Risk Dataset:**
  - Can show an individual the tree and have them easily follow if they have or do not have the symptoms on the tree to show their chance of having a stroke.

RANDOM FOREST

- An ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting.
- Works by averaging the predictions of multiple decision trees trained on different subsets of data.
- More robust than a single decision tree, reducing variance while maintaining interpretability – not as easy to visualize as a decision tree.
- **Why used?**
  - Provides higher accuracy and stability compared to a single decision tree.
  - Reduces overfitting by averaging multiple models, leading to better generalization.
  - Handles nonlinear relationships effectively.
- **Relevance to Stroke Risk Dataset:**
  - Can model complex interactions (ex. the multiple factors in this data set) without requiring explicit assumptions.

KNN (K-Nearest Neighbors)

- A supervised learning technique that predicts the response by getting the response for the k nearest neighbors.
- Works by averaging the predictions of the k nearest neighbors, which are found by a search algorithm using a distance metric, which here is Euclidean Distance.
- Different k values are tested using cross-validation to find the optimal one.
- Requires normalization of predictors beforehand to prevent high magnitude factors unduly influencing the results.
- **Why did we use it here?**
  - Similar patients are likely to have similar stroke risks.
  - Simplicity of model makes it easier to explain to medical professionals.

Gradient Boosting

- An ensemble learning method that builds multiple decision trees sequentially, where each tree corrects the errors of the previous one, thereby improving accuracy and performance.
- Works by building trees iteratively, focusing on reducing errors from prior models.
- Like Random Forest, more robust than a single decision tree, reducing variance while maintaining interpretability, but not as easy to visualize as a decision tree.
- **Why used?**
  - Provides high accuracy by minimizing errors iteratively.
  - Reduces bias while maintaining low variance.
  - Can work well with complex, nonlinear relationships.

Naïve Bayes

- Naïve Bayes is a probabilistic Machine Learning Model that assumes all of our variables are independent.
- The model assumes all variables are independent, which provides a simpler method than other models.
- The assumption of all variables being independent provides an advantage of finding unknown relationships between health metrics and potential stroke that can be used for analysis in identifying need for medical intervention.
- **Why used?**
  - Simpler model by having independent probabilities for each variable, helping with explainability of results.
  - May uncover unknown relationships between our variables and potential of stroke.

Beta Regression

- Beta Regression is similar to a linear model; however, it uses a continuous distribution for our y.
- We will use all variables for this model (Except for “At Risk”) and solve for the continuous variable of “Stroke Risk Percentage” (Distributed between 0-100%). We’ll divide these by 100, as they need to be put on a 0-1 scale, and, we will have to subtract .000001 from these amounts, as any “0” or “1” breaks our model.
- Beta Regression can handle asymmetry and heteroskedasticity as it is a non-linear model.
- Some variables that may not be significant in linear models may be significant in Beta Regression.
- **Why used?**
  - Handles potential asymmetry and heteroskedasticity
  - Fit non-linear relationships
  - Can use the continuous response variable for stroke risk.
  - Potentially identify other features that are significant with this non-linear, continuous distribution model.

FEATURE SELECTION

- Most methods performed feature selection implicitly.
- RFE was used in the case of logistic regression.
- Based on our rankings, we chose the **bolded** features for our final models.
- Using all features (with age) gave us model accuracies between 90-100% which is not realistic. Hence Age was removed.
- Given at triage, patients might not have all symptoms, we want to choose the ‘best’ possible stroke risk predictors to have respectable results.

	Feature Rank for each method								
Feature	**Logistic Regression**	**Ridge Regression**	**LASSO**	**Elastic Net**	**Naïve Bayes**	**Decision Tree**	**Random Forest**	**xgBoost**	**Beta Regression**
Cold Hands Feet	1	1	1	1	9	2/3 & 4	1	1	1
Persistent Cough	8	8	8	8				7	
Fatigue Weakness	7	7	7	7	3		9	9	9
Chest Pain	9	10	9	9	1	1	4	2	2
Excessive Sweating	5	5	5	5	8		3	3	3
High Blood Pressure	4	4	4	4	5			4	4
Anxiety feeling of Doom	2	2	2	2	4	2/3	10	5	6
Nausea Vomiting	6	6	6	6			6	6	5
Sleep Apnea	3	3	3	3	2		8	7	7
Dizziness					6				
Irregular Heartbeat	10	9	10	10	7	2/3			8
Shortness of Breath							5	10	
Swelling Edema							7	8	

RESULT AND NEXT STEPS

- Age was the most significant factor of stroke risk, so much so that it causes overfitting of our models.

- We ran our models without the "Age" variable to determine which variables were significant in each of our models for variable selection.

- We determined the top features were:
  - **Cold Hands Feet, Fatigue Weakness, Chest Pain, Excessive Sweating, Anxiety feeling of Doom, Nausea Vomiting, and Sleep Apnea -** Consistently ranked highly across multiple models.
  - **High Blood Pressure -** Major stroke risk factor in medical literature + moderate rankings (Rank 6 in multiple models).
  - **Irregular Heartbeat –** Clinically essential predictor (even if not ranked highly by models, it’s a major stroke risk factor).

NEXT STEPS

- Run and obtain model results for all models using the top 8 variables.
  - Then run the models again adding the variable "Irregular Heart Rate" (total of 9 variables) to assess if there is a significant change in performance.

- Compare and analyze model results
  - Determine which model had the most accurate results overall.
  - Determine which model could be used for a questionnaire to triage patients with potential stroke risk, considering Sensitivity and Specificity for over/under classifying positive stroke risk cases.
  - Compare traditional methods (Regression) and Machine Learning methods (Random Forest, SV) for differences in results.

PROBLEM STATEMENT (WHAT HAVE WE LEARNT SO FAR)

- Can we create a questionnaire for patient intake to appropriately triage patients for stroke? - perhaps, as not all risk factors would be available to medical staff in the case of an episode.
- Can we create an accurate model for predicting stroke risk assuming precise medical measurements have been done? - in progress, we hope so!
- What are the most significant risk factors associated with a chance of having a stroke? – Apart from Age, we are trying to answer this question.
- How does age impact the likelihood of experiencing a stroke? – The likelihood of stroke increases as Age increases.
- Are there any significant interactions between risk factors predicting stroke? - Given the synthetic nature of the dataset, the predictors, though independent, seem to come together to predict stroke risk very well.
- Traditional method (regression) versus ML (random forest, SVM) differences? - in progress.

THANK YOU

Stay tuned for the full report!