Executive Summary

Group 3: Chris Cao, Daoud Haq, Brian Moritz, Luke Natalo

Applying for life insurance is a gruesome and lengthy process. We want to explore if we can use machine learning to help expedite the process by pre-determining a candidate's risk level according to a few inputs provided by the customer. This would in turn reduce costs for the insurance company. We've obtained a dataset from Prudential containing approximately 60,000 rows of anonymous customer data containing column fields such as: current life insurance policy, employment history, family history, height, weight, BMI, and lots of medical history questions. Due to HIPAA, these columns are coded as "Employment Question 1," "Employment Question 2," "Family History Question 1," etc. and the values are entered as zero for "no" or a one for "yes." We will use this data to create multiple classification models and the one with the highest accuracy will be chosen.

Below is a list of the questions that we will be asking and answering:

1. What type of model should be used (e.g., classification or regression?)
2. Which specific model would work best for the type of data given (after determining the model type to use)?
3. How accurate can we make the model and how (e.g., which parameters to tune and by how much)?
4. What other parts of underwriting can we automate?
5. Can we also use ML to predict the outcome of a claim assuming it is legitimate?
6. How much cheaper or efficient can using ML be rather than traditional under writing?
7. How much money do life insurance companies spend on administrative costs per year?