# Capstone Two: Project Proposal

Prudential Life Insurance Assessment Dataset

*By: Brianna Abdalla*

## Problem:

What opportunities exist for Prudential Life Insurance to shorten the application process and assess the risk of new applicants with 95% or greater accuracy in the next year?

## Description:

Prudential Life Insurance is one of the largest issuers of life insurance in the United States. In the past, life insurance applications were a long and tedious process for both customers and insurance agents. A traditional application requires new customers to devote a significant amount of time and effort in the process. Application requirements include a mound of paperwork, detailed medical history and even a blood draw from the potential customer. This traditional application method can take up to 30 days to process! In the modern era where consumers expect results instantaneously, a 30 day processing speed presents a significant issue for life insurance companies resulting in only 40% of U.S. households with individual life insurance.

Prudential Life Insurance wants to help people protect their families by creating an online risk assessment that will not only speed up the application process, but predict an applicant's risk with 95% or greater accuracy. With this goal in mind, Prudential provided a dataset on the Kaggle website that contains 59,381 insurance applications with their resulting risk score. The idea is to filter applications based on health information responses and correctly predict the associated risk score. If a successful model can be created, Prudential can provide the application instantly to customers from an online platform and significantly increase the application process speed times.

## Criteria for Success:

Over the next month, the Prudential Life Insurance dataset will be cleaned, organized and modeled to predict the risk factor (on a scale from 1-8) based on the given application values. This model will then be applied to a test set and adjusted until a 95% or greater accuracy is achieved.

## Scope of Solution Space:

The Prudential Life Insurance Dataset Project will focus only on the given Kaggle dataset. Once successful risk prediction is achieved, Prudential can choose to implement the model into their application procedure and achieve faster processing times.

## Constraints within Solution Space:

- The dataset has over 100 heath indicator columns all with ambiguous titles and continuous values making it very difficult to draw any inference from them without modeling.
- Even if the dataset is modeled with 95% accuracy, it is possible for future applications to differ significantly from the test set and be incorrectly categorized.
- There are many health indicators that have null values indicating that an applicant does not have that particular health issue. The majority of the null values cannot be dismissed because they could significantly impact the risk assessment.
- We have to assume that the given life insurance applications were assessed accurately and provide true risk evaluations.

Stakeholders:

- Prudential Life Insurance CEO and management team
- Prudential Data Science and Engineering team

Key Data Sources:

CSV File: train.csv - This is the dataset that will be used to clean and train the data.

CSV File: test.csv - This is the provided test set that our model will be used on.

https://www.kaggle.com/competitions/prudential-life-insurance-assessment/data

Problem Approach Plan:

1. After loading the dataset, I will complete the data wrangling phase including cleaning and organizing the data. I would also like to consider how to treat the null values and determine if a different place holder is necessary for them.
2. I will begin to explore how the health indicators influence the risk score and begin searching for correlations. I will probably start looking at regression analysis here.
3. A test set has been provided for the pre-processing phase. I will screen out any out of value ranges here and prepare the data for modeling.
4. Lastly, I will model the data. I am assuming that the model will be linear.

Deliverables:

- GitHub repository containing work completed for each step
- Presentation slide deck
- Project Final Report