Project Proposal:
Predicting Risk for Life Insurance Applicants

Springboard Data Science Track: Independent Capstone 2 Project

By: Brianna Abdalla

Problem:

What opportunities exist for Prudential Life Insurance to identify the attributes associated with risk, shorten the application process and assess the risk of new applicants within 95% confidence?

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 (at the time of the dataset release date) 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 accuracy that can be supported within 95% confidence. 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 classify applications based on health information responses and correctly predict the associated ordinal risk score. If a successful model can be created, Prudential can use it to significantly speed up application processing times and improve new customer acquisition.

Criteria for Success:

The Prudential Life Insurance dataset will be cleaned and organized with relevant relationships identified. This data will then be split into testing data and evaluated on AUC and accuracy scores then cross validated to identify a model that preforms within the 95% confidence interval.

Scope of Solution Space:

The 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 120 variables that were normalized with ambiguous titles, making them difficult to interpret intuitively.
- The 8 ordinal classes are heavily imbalanced with around 30% of the data falling in class 8.
- There is significant missing data in several features, yet some of these features seem to relate to the target variable and completely removing them is not an option.

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.
- Link to Kaggle competition: 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.
3. I will create testing data by splitting the train data.
4. Lastly, I will model the data and select the best model based on AUC and accuracy.

Deliverables:

- PDF Project Proposal
- GitHub repository:
  - Notebooks for wrangling, EDA, preprocessing, and modeling.
  - A file called Models for saved top models
  - A file called Model Metrics for saved best model metrics
- Presentation slide deck
- PDF Project Final Report