Machine Learning Engineer Nanodegree

Capstone Proposal

Hamdy Tawfeek

January 17th, 2019

Proposal

Domain Background

In a one-click shopping world with on-demand everything, the life insurance application process is antiquated. Customers provide extensive information to identify risk classification and eligibility, including scheduling medical exams, a process that takes an average of 30 days.

The result? People are turned off. That’s why only 40% of U.S. households own individual life insurance. Prudential wants to make it quicker and less labor intensive for new and existing customers to get a quote while maintaining privacy boundaries.

Problem Statement

Can you make buying life insurance easier?

By developing a predictive model that accurately classifies risk using a more automated approach, you can greatly impact public perception of the industry.

Datasets and Inputs

The dataset provided by Prudential on Kaggle.
In this dataset, a hundred variables describing attributes of life insurance applicants. The task is to predict the “Response” variable for each Id in the test set. “Response” is an ordinal measure of risk that has 8 levels.

File descriptions

- train.csv - the training set, contains the Response values
- test.csv - the test set, you must predict the Response variable for all rows in this file

Data fields

Variable	Description
Id	A unique identifier associated with an application.
Product_Info_1-7	A set of normalized variables relating to the product applied for.
Ins_Age	Normalized age of applicant.
Ht	Normalized height of applicant
Wt	Normalized weight of applicant

BMI	Normalized BMI of applicant
Employment_Info_1-6	A set of normalized variables relating to the employment history of the applicant.
InsuredInfo_1-6	A set of normalized variables providing information about the applicant.
Insurance_History_1-9	A set of normalized variables relating to the insurance history of the applicant.
Family_Hist_1-5	A set of normalized variables relating to the family history of the applicant.
Medical_History_1-41	A set of normalized variables relating to the medical history of the applicant.
Response	This is the target variable, an ordinal variable relating to the final decision associated with an application

The following variables are all categorical (nominal):

Product_Info_1, Product_Info_2, Product_Info_3, Product_Info_5, Product_Info_6, Product_Info_7, Employment_Info_2, Employment_Info_3, Employment_Info_5, InsuredInfo_1, InsuredInfo_2, InsuredInfo_3, InsuredInfo_4, InsuredInfo_5, InsuredInfo_6, InsuredInfo_7, Insurance_History_1, Insurance_History_2, Insurance_History_3, Insurance_History_4, Insurance_History_7, Insurance_History_8, Insurance_History_9, Family_Hist_1, Medical_History_2, Medical_History_3, Medical_History_4, Medical_History_5, Medical_History_6, Medical_History_7, Medical_History_8, Medical_History_9, Medical_History_11, Medical_History_12, Medical_History_13, Medical_History_14,

Medical_History_16, Medical_History_17, Medical_History_18, Medical_History_19, Medical_History_20, Medical_History_21, Medical_History_22, Medical_History_23, Medical_History_25, Medical_History_26, Medical_History_27, Medical_History_28, Medical_History_29, Medical_History_30, Medical_History_31, Medical_History_33, Medical_History_34, Medical_History_35, Medical_History_36, Medical_History_37, Medical_History_38, Medical_History_39, Medical_History_40, Medical_History_41

The following variables are continuous:
Product_Info_4, Ins_Age, Ht, Wt, BMI, Employment_Info_1, Employment_Info_4, Employment_Info_6, Insurance_History_5, Family_Hist_2, Family_Hist_3, Family_Hist_4, Family_Hist_5

The following variables are discrete:
Medical_History_1, Medical_History_10, Medical_History_15, Medical_History_24, Medical_History_32
Medical_Keyword_1-48 are dummy variables.

Solution Statement

In this project, I would like to use compare different predictive classification models and see which is better. I will implement logistic regression, XGBoost, Neural Network.

Benchmark Model

Prudential Kaggle competition used a benchmark score of 0. So, I will

take a score of 0 as my benchmark for this project.

Evaluation Metrics

Submissions are scored based on the quadratic weighted kappa, which measures the agreement between two ratings. This metric typically varies from 0 (random agreement) to 1 (complete agreement). In the event that there is less agreement between the raters than expected by chance, this metric may go below 0.

The response variable has 8 possible ratings. Each application is characterized by a tuple (ea,eb), which corresponds to its scores by Rater A (actual risk) and Rater B (predicted risk). The quadratic weighted kappa is calculated as follows.

First, an N x N histogram matrix O is constructed, such that Oi,j corresponds to the number of applications that received a rating i by A and a rating j by B. An N-by-N matrix of weights, w, is calculated based on the difference between raters' scores:

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

An N-by-N histogram matrix of expected ratings, E, is calculated, assuming that there is no correlation between rating scores. This is calculated as the outer product between each rater's histogram vector of ratings, normalized such that E and O have the same sum.

From these three matrices, the quadratic weighted kappa is calculated

as:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}.$$

Project Design

The project will roughly consist of the following steps:

1. Understanding the data.
2. Data Preprocessing
3. Model Building and evaluation.
4. Compare different results from different models and choose the model with the highest predictive power.