
JOB-A-THON

# The Complete Approach

Name: Srihari R

Mail Id: sriharikrishna06@gmail.com

Private Rank: 271

**Problem Statement:**

Your Client FinMan is a financial services company that provides various financial services like loan, investment funds, insurance etc. to its customers. FinMan wishes to cross-sell health insurance to the existing customers who may or may not hold insurance policies with the company. The company recommend health insurance to its customers based on their profile once these customers land on the website. Customers might browse the recommended health insurance policy and consequently fill up a form to apply. When these customers fill-up the form, their Response towards the policy is considered positive and they are classified as a lead.

Once these leads are acquired, the sales advisors approach them to convert and thus the company can sell proposed health insurance to these leads in a more efficient manner.

Now the company needs your help in building a model to predict whether the person will be interested in their proposed Health plan/policy given the information about:

1. Demographics (city, age, region etc.)
2. Information regarding holding policies of the customer
3. Recommended Policy Information

**Data Information:**

The training data consists of 50882 and the test data consists of 21805.

There are 13 features of these data which are **'City_Code', 'Region_Code', 'Accomodation_Type', 'Reco_Insurance_Type', 'Upper_Age', 'Lower_Age', 'Is_Spouse', 'Health Indicator', 'Holding_Policy_Duration', 'Holding_Policy_Type', 'Reco_Policy_Cat', 'Reco_Policy_Premium', 'Response'.**

The target/dependent feature is **'Response'**

**Data Cleaning and Preparation:**

Nulls values are found in 'Health Indicator', 'Holding_Policy_duration' and 'Holding_policy_type'. The nulls values in these features are imputed with mode values for '**Health Indicator**', '**Holding_Policy_Type**' and null values in '**Holding_Policy_Duration**' are imputed with its median value.

**'Accomodation_Type', 'Reco_Insurance_Type'** and **'Is_Spouse'** are label encoded and transformed to numerical types.

**'Holding_Policy_Duration', 'Reco_Policy_Premium', 'Upper_Age', 'Lower_Age'** are scaled using Standard Scaled, these values scaled in such a way that its mean value will be approximately zero.

**Data Modelling**

The Final features chosen are 'Region_Code', 'City_Code', 'Upper_Age', 'Is_Spouse', 'Reco_Policy_Cat', 'Holding_Policy_Duration' based on manual selection.

KNN and Decision Tree is modelled and the best model is selected based on the roc score. The data given to us for modelling is split into train and test dataset with 70:30 ratio.

KNN is used first for modelling, its hyperparameters are,

1. Algorithm: KD Tree
2. Weights: Distance
3. N neighbors: 2

The train score is 0.9999 and its test score is 0.6300. This is showing overfitting.

Decision Tree is modelled using the train set and roc score is 0.9433 and the test score is 0.59433. This is also clearly showing signs of overfitting.

Since both the algorithms are showing overfitting, we had to choose the best among them and chose decision tree since it showed less overfitting among the two.

In order to get best decision tree model, bagging classifier is used with 100 estimators i.e., the model runs 100 times and the best among them is chosen. The best model's roc score is 0.9973 and test score is 0.6779

**Testing on Test data provided**

The test data provided for us went through same process for data cleaning and preparation as train data.

The final probability values of the predicted values are submitted along with the ID.

Upon submission this algorithm received a roc score of 0.6771. The **Private Rank awarded was 271.**

**Areas to Improve in future**

1. Improve in dealing with categorical features
2. Improve in feature selection
3. Improve in hyperparameter tuning