AlmaBetter

Health Insurance Cross Sell Prediction

(Technical Documentation)

BY – AYUSH SHARMA

PROBLEM STATEMENT

Our client is an insurance company that has provided Health Insurance to its customers now they need your help in building a model to predict whether the policyholders (customers) from past year will also be interested in Vehicle Insurance provided by the company.

DATA OVERVIEW

The dataset provided consisted of 381109 rows and 12 columns.

we are given the following columns in our data

1. **id: Unique ID for the customer**

2. **Gender: Gender of the customer**

3. **Age: Age of the customer**

4. **Driving License 0: Customer does not have DL, 1: Customer already has DL**

5. **Region Code: Unique code for the region of the customer**

6. **Previously Insured: 1: Customer already has Vehicle Insurance, 0: Customer doesn't have Vehicle Insurance**

7. **Vehicle Age: Age of the Vehicle**

8. **Vehicle Damage:1: Customer got his/her vehicle damaged in the past. 0: Customer didn't get his/her vehicle damaged in the past.**

9. **Annual Premium: The amount customer needs to pay as premium in the year**

10. **Policy Sales Channel: Anonymized Code for the channel of outreaching to the customer i.e. Different Agents, Over Mail, Over Phone, In Person, etc.**

11. **Vintage: Number of Days, Customer has been associated with the company**

12. **Response: 1: Customer is interested, 0: Customer is not interested**

Performing EDA (exploratory data analysis)

A. Exploring head and tail of the data to get insights on the given data.

B. Looking for null values and removing them if it affects the performance of the model.

C. Exploring descriptive summary of the dataset.

D. Plotting Columns like Response, Vehicle Damage, Gender, Driving License, Vehicle Age.

E. Plotting Response with respect to gender, vehicle damage, previously insured, vehicle age.

F. Encoding the categorical columns for better performance of our model.

MODELS USED

- ### RANDOM FOREST -:

We first used Random Forest model and fit it on Training set. Then we test the model on test set and evaluated the model performance by using evaluation metrics
Accuracy – 0.8327
Precision– 0.7798
Recall– 0.926
Roc Auc score- 0.832

- ## LOGISTIC REGRESSION -:

  We then used Logistic Regression model and fit it on Training set. Then we test the model on test set and evaluated the model performance by using evaluation metrics

  Accuracy – 0.794

  Precision– 0.742

  Recall– 0.900

  Roc Auc score- 0.794

- ## DECISION TREE -:

  We then used Decision Tree model and fit it on Training set. Then we test the model on test set and evaluated the model performance by using evaluation metrics

  Accuracy – 0.807

  Precision– 0.775

  Recall– 0.864

  Roc Auc score- 0.807

- ## K-Nearest Neighbors -:

  We then used K-Nearest Neighbors model and fit it on Training set. Then we test the model on test set and evaluated the model performance by using evaluation metrics

  Accuracy – 0.81

  Precision– 0.74

  Recall– 0.95

  Roc Auc score- 0.81

- ## NAIVE BAYES -:

  At last, we used Naïve Bayes model and fit it on Training set.

Then we test the model on test set and evaluated the model performance by using evaluation metrics
Accuracy – 0.79
Precision– 0.73
Recall- 0.91
Roc Auc score- 0.79

conclusion

In the conclusion I want say that from the beginning we did data inspection. After that we perform EDA and were able to draw relevant conclusions from the given data and then we trained our model on Random Forest and other models.

Out of all models used, with Random Forest model we were able to get the highest accuracy, precision, roc auc score of 0.832, 0.779, 0.832 respectively.