Capstone Project: Insurance Company Benchmark

Manohari Wijesooriya

Student ID: 501212269

The Chang School of Continuing Education, Toronto Metropolitan University

CIND820 Big Data Analytics Project

Dr. Tamer Abdou

May 15, 2023

Contents

1. Abstract

This data set is about a direct marketing case from the insurance sector which was to predict and explain policy ownership. It is about predicting who would be interested in buying a caravan insurance policy and to give a relevant explanation. If the company had a better understanding of who their potential customers were, they would know more accurately who to send policy quotes to, so some of this waste and expense could be reduced.

Problem statement:

Can you predict who would be interested in buying a caravan insurance policy and give an explanation why?

Following tasks will be performed in this analysis.

- Predict which customers are potentially interested in a caravan insurance policy.
- Describe the actual or potential customers; and possibly explain why these customers buy a caravan policy.

Dataset: **Insurance Company Benchmark (COIL 2000).** This data set used in the CoIL 2000 Challenge contains information on customers of an insurance company. The data consists of 86 variables and includes product usage data and socio-demographic data

Dataset can be found in this link : https://archive.ics.uci.edu/ml/datasets/Insurance+Company+Benchmark+%28COIL+2000%29

TICDATA2000.txt: Dataset to train and validate prediction models and build a description (5822 customer records). Each record consists of 86 attributes, containing sociodemographic data (attribute 1-43) and product ownership (attributes 44-86).The sociodemographic data is derived from zip codes. All customers living in areas with the same zip code have the same sociodemographic attributes. Attribute 86, "CARAVAN: Number of mobile home policies", is the target variable.

TICEVAL2000.txt: Dataset for predictions (4000 customer records). It has the same format as TICDATA2000.txt, only the target is missing. Participants are supposed to return the list of predicted targets only. All datasets are in tab delimited format. The meaning of the attributes and attribute values is given below.

TICTGTS2000.txt Targets for the evaluation set.

The research is to predict whether a customer is interested in a caravan insurance policy from other data about the customer. In this research classification and regression analysis will be used for the prediction part. First we will model using several algorithms; Naïve base,

sklearn, Decision Tree, Random Forest, Balanced Random Forest, logistic regression and kneighbors. Then results of these models will be compared (accuracy, precision and recall will be examined) and select the best algorithm to predict caravan insurance customer.

The purpose of the description task is to give a clear insight to why customers have a caravan insurance policy and how these customers are different from other customers. Descriptions can be derived using regression equations and decision trees. In this task rules will be set to identify Caravan insurance customer and measure statistical significance of those rules.

Python will be used to perform analysis and Google Collab will be used to run Python script. Majority of time in this research will be spent on cleaning and understanding of data variables. Tableau will be used in profiling dataset and visualizations.