**DATS 6202 Machine Learning I**

*Final Project Proposal*

Group 5: Lianjie Shan, Xi Zhang

**Problem Selection and Research Objectives**

**Description of the Question**

We based our research on the subject of car insurance. The main problem is classifying customers by characteristics for insurers. The target is to provide more information to the car insurance market and make transactions more viable and efficient.

**Research Overview**

Because the content of automobile insurance needs to be considered is too complex, and the relationship between many variables cannot be directly and simply judged, we hope to train several models through machine learning. Our goal is to accurately judge whether the insurer will make a claim and the extent of the claim after the insurance is purchased from the insurance company through the personal information, family background, asset status, historical insurance records, and the insured vehicles. We can classify different insured persons and guide insurance companies. So that the insurance company can quickly determine the type of insured customers and claims probability, and thus avoid the loss of the insurance company. It can also provide customers with more targeted insurance plans.

**Related Material and Background Supportive**

To ensure that our ideas have a solid theoretical foundation and prove to be feasible, we looked at researches on deep learning in the insurance industry. In Sato’s research, he pointed that “AXA, the large global insurance company, has used machine learning in a POC to optimize pricing by predicting “large-loss” traffic accidents with 78% accuracy” (Kaz Sato, 2017). This proves that in recent years the insurance industry has been using machine learning to optimize prices. He takes age range of the driver, region of the driver's address, annual insurance premium range and age range of the car as outputs to train the multilayer neural network model (Kaz Sato, 2017). I believe this practice can also implement our research.

By *MACHINE LEARNING IN INSURANCE*, we can also learn from that machine learning is changing the way insurance companies interact with customers. “Consumers are seeking personalized solutions—made possible by machine learning algorithms that review their profiles and recommend tailor-made products” (Ravi Malhotra, Swati Sharma, 2018). Malhotra and Sharma illustrate the huge value of machine learning in insurance and give examples of how it optimizes market matching. This is what our research is trying to achieve.

In the book *Python Machine Learning*, it provides most of the technical support. From theory to finding the right model and converting mathematics to the python language. This book will serve as a guide for our research.

**Data Selection and Overview**

**Data Source**

The data we chose was released by Kaggle, an open-source data site. The distributor xiaomengsun published it in 2018. It is made up of a record of 3,092,684 observations and 27 variables. This data can be downloaded from the following websites for study and research:

<https://www.kaggle.com/xiaomengsun/car-insurance-claim-data>

**Data Views:**

|  |  |  |  |
| --- | --- | --- | --- |
| # | Columns | Description | Data type |
| 1 | ID | Customers’ ID Number | int |
| 2 | KIDSDRIV | The number of kids the customers need to drive for. | int |
| 3 | BIRTH | Date of birth. | Factor |
| 4 | AGE | Age | int |
| 5 | HOMEKIDS | The number of kids the customers have. | int |
| 6 | YOJ | Working age. | int |
| 7 | INCOME | Income. | int |
| 8 | PARENT1 | Whether the customers’ parents pay for their insurance. | Factor |
| 9 | HOME\_VAL | Property value. | Factor |
| 10 | MSTATUS | Marriage status. | Factor |
| 11 | GENDER | Gender. | Factor |
| 12 | EDUCATION | Education level. | Factor |
| 13 | OCCUPATION | Occupation. | Factor |
| 14 | TRAVTIME | Travel times. | int |
| 15 | CAR\_USE | Whether customers’ cars are used for private or commercial. | Factor |
| 16 | BLUEBOOK | Blue Book is a guidebook that compiles and quotes prices for new and used automobiles and other vehicles of all makes, models and types. | Factor |
| 17 | TIF | The insured vehicle. | int |
| 18 | CAR\_TYPE | Car Type. | Factor |
| 19 | RED\_CAR | Red cars are more expensive to insure. | Factor |
| 20 | OLDCALIM | Cumulative claim amount. | Factor |
| 21 | CLM\_FREQ | Claim frequency. | int |
| 22 | REVOKED | Revoked time. | Factor |
| 23 | MVR\_PTS | Motor vehicle record points. The Violation/Accident Guidelines and Points Columns on the right are used to assign points to each accident or violation over a three-year period. | int |
| 24 | CLM\_AMT | Claim amount. | Factor |
| 25 | CAR\_AGE | Car age. | int |
| 26 | CLAIM\_FLAG | Claim flag. | int |
| 27 | URBANICITY | Car insurance varies according to the area of activity. | Factor |

**Initial Analysis and Procedure for Project**

**Data Preprocessing Section**

To begin with, we will examine data to judge whether there exist missing values in a dataset. At the same time, we also test if there are outliers in the data and delete them from the original dataset. Since some features in our dataset are similar and have no meaning with our target, we will delete some features(columns) to reduce the dimension. After finishing the steps above, we will get a completer and more manageable dataset.

**Target Division and Feature Selection Section**

For target division, we will divide our target into four classes, which are no claim class, less claim class, medium claim class, and high claim class. And then, we will convert the value of four categories from continuous data to discrete data. In this way, we can provide insurance companies with a relatively accurate delineation of the insured, which has certain reference value for insurance companies to provide different insurance plans for different groups.

For the feature selection, we test the correlations between the target and other features in the dataset. From this, we can filter out some insignificant parameters and further reducing dataset. And then, we can test the relationships between the rest of the features and target.

**Classifier Selection and Framework Design Section**

After data preprocessing, we split our dataset into a training set (70% of data) and testing set (30% of data). For the training set, we will apply the neural network and backpropagation algorithm to build a predictive model and use boosting technic to improve our results and model. For the testing set, we use our model to predict the outcome of the testing data to determine the feasibility of the model. Besides, we plan to set up four neurons and obtain four outputs.

**Evaluation Section**

To test the accuracy of the predictive model we built, we will use ROC and AOC methods. And then, we can analyze the feasibility of the model.

**Summary and Further Development of the Project Section**

From our final results and model, we can predict the relative size of the probability of needing a claim in the future based on the characteristics of the insured person and the car, which is helpful for the insurance companies. To improve our model and practical value, we can divide target into more categories and select more classifiers to provide a more detailed and accurate classification for insurance companies.

**Schedule for Finishing the Project**

|  |  |
| --- | --- |
| **Schedule** | **goals** |
| July 25th – August 1st | Data processing section and Target division and feature selection section |
| August 2nd – 9th | Classifier selection and framework design section |
| August 10nd – 16th | Evaluation section |
| August 17nd – 20th | Summary and Presentation section |

**References**

Sato, Kaz. “*Using Machine Learning for Insurance Pricing Optimization* | Google Cloud Blog.” *Google*, Google Cloud Platform, 19 Mar. 2017, cloud.google.com/blog/big-data/2017/03/using-machine-learning-for-insurance-pricing-optimization.

Malhotra, Ravi, and Swati Sharma. *MACHINE LEARNING IN INSURANCE - Accenture.com*. Accenture, 2018, www.accenture.com/t20180822T093440Z\_\_w\_\_/us-en/\_acnmedia/PDF-84/Accenture-Machine-Leaning-Insurance.pdf.

Raschka, Sebastian, and Vahid Mirjalili. *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow*. Pack Publishing, 2018.