Santosh Tirunagari
245577
Santosh.tirunagari@aalto.fi

Impact of other policies and automobiles on caravan policy

Abstract:

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 derived from zip area codes. This report gives the visualization for the trends and interesting features of the dataset for the given hypothesis.

Hypothesis:

- ✓ Persons having caravan policy will have a high probability of having a car, because the caravan must be carried by an automobile (Car is cheap and easy medium of transport for caravan).
- ✓ Hence they will also have a car policy and fire policy (caravans have high risk of catching fire because of the materials used in constructing).
- ✓ They also will have a boat policy because boating and caravanning are indicative of an outdoor lifestyle.
- ✓ Caravan policy will be most taken by low Status people, who use caravans as mobile homes and high status people for outdoor lifestyle.

Introduction:

"One picture is more valuable then thousands of words". It takes much time to analyze data in a table, but couple of seconds to analyze visualization. Huge amounts of information available today made us deal on information visualization. It supports the decision making in various departments. Investigating and analyzing large amounts of data is seriously a difficult task, but by using some data mining or machine learning techniques through information visualization can make it rather simple.

Data preprocessing and Feature Selection:

The CoIL 2000 dataset contains 5822 samples or instances of customers with 86 features. Where 86^{th} feature gives the binary information about a customer having the caravan policy or not. Hence the COIL data set encloses 85 possible, input features. I cannot rely on all the variables; hence needed to find a reliable subset of features for making the hypothesis which is the first important task. I decided to write a simple script in matlab which can give me the counts of each feature who are having a caravan policy. Hence using the high count of the features I have the stated above hypothesis.

The method of finding a subset of features is called feature selection. I have selected the following features for making the hypothesis. 1) Car policies 2) Fire policies 3) Social Status 4) Boat policies. I also have selected some other related features to compare them with the selected above mentioned features.

Methods of Visualization:

Doughnut chart: A doughnut chart is a kind of Pie chart with a blank at the center; Doughnuts have an ability to support multiple statistics as one. A simple pie chart is typically used to show the proportion of data within a single category. Rather than using a doughnut chart to increase the number of categories that can be displayed in a single chart, a doughnut chart may be used to show greater levels of details across a single category of information. It displays the contribution of each variable to the total.

Horizontal Bar chart: Horizontal Bar charts are used to compares the values across the categories using the horizontal bars or rectangles. It is used when the categories represent durations or when the categories names are too long to represent the data.

Column chart: Column charts are also called as bar charts, these are used to compares the values across the categories using the vertical rectangles. It is used when the order of the categories is not very considered or when displaying the category counts.

Experiments: According to the stated hypothesis cars are used mostly to carry the caravans as they are cheap and efficient means of transport. Hence I used the contributions of the different automobiles policies which influence the caravan policy. This visualization is done using doughnut chart. It is clear from the Figure1; that there are 8 categories or classes for automobiles and the contribution of Car is larger than the rest of the categories of

Santosh Tirunagari
245577
Santosh.tirunagari@aalto.fi

automobiles. It could have been more aesthetic to use the 3D visualization for the doughnut charts, but it would increase the data-ink ratio and also considered to be the chart junk. Hence this visualization is based on the principles of Tufte. Multiple colors used in the chart clearly discriminate each category with the quantity of its contribution.

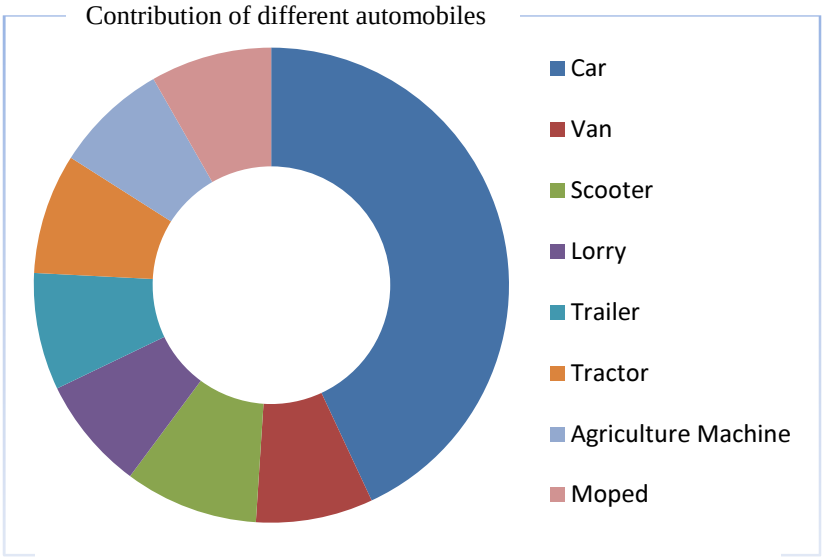

Figure 1: Contribution of different automobiles

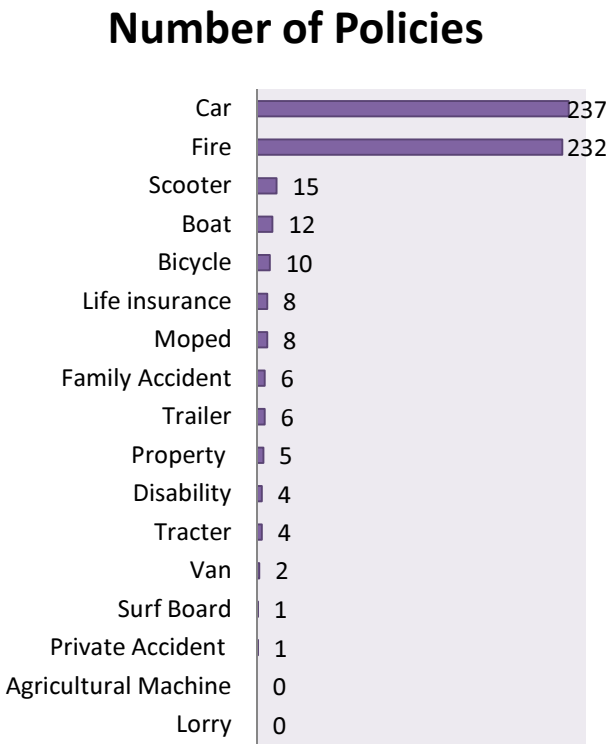

Figure 2: Counts of Number of Policies.

As caravans are made of such a kind of material which is prone for fire accidents, the fire policies would also influence the caravan policy. From the Figure1, as the contribution of the car automobile is more, it is obvious that the car would have a car policy. Hence car policy would also influence for a caravan policy. The Figure2 shows us clearly that the fire policy and the car policy have most number of hits for having a caravan policy.

But according to the hypothesis customers having a boat policy would also have a caravan policy because boating and caravanning is an outdoor lifestyle of living. From the figure2 it is clear that the hits for the boat policy are very less. This is because of the low status people who are using caravans only for the living purpose as their mobile homes, but not for some outdoor lifestyles. I run the matlab script to find number of low status customers having boat policy and found it was only 9.

In figure2 the data is visualized using a horizontal bar chart. The visualization is truly based on Tufte principles.

Santosh Tirunagari
245577
Santosh.tirunagari@aalto.fi

Impact of other policies and automobiles on caravan policy

From the Figure2, we can also see that the policies of automobiles like Lorries and Agricultural machines have no hits for the caravan policy. This is may be because the lorry policy customers have their own transportation system and generally they are used for trading and carrying goods. The agricultural machine policy customers are more focused on the agricultural land for their cultivations and they do not show much interest in moving from one place to another.

There is no need of multiple colors for the visualization because they belong to same category and more over there is no other feature discriminating these categories. Single color is used in the other figures only to show that they belong to same category which obeys the gestalt's laws.

Figure3 visualizes the features that were hidden in the figure2. It was clear in the figure2 that all the features except car and fire were hidden because of the high hits of the car and fire policies. After removing the car and fire policies we can visualize the trends in the hidden content of figure2 more properly.

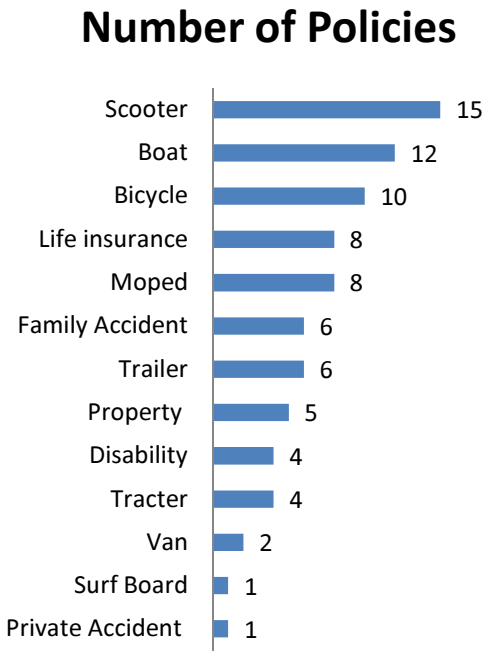

Figure3: Visualization of Hidden features in figure2.

From the hypothesis, I stated that the caravan policies are taken mostly by a low status customers, using caravans as their mobile homes and high status customers for their outdoor lifestyles. From the Figure4, it's clearly seen that the social status A, B1 (High Status) customers have more hits for caravan policies and social status D (Low status) customers have high hits for the caravan policies.

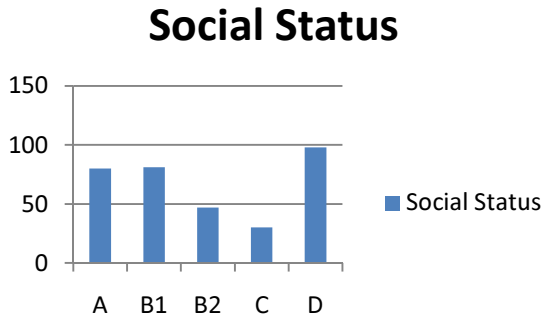

Figure4: Social status counts for caravan policy holders

In Figure4, It is clear that there are 5 categories or classes of social status in the data. I have purposely added unnecessary grid lines and the color indicator 'social status' on the right hand side so that we could see an increased DI-ratio and I also avoided it, using Tufte's principles, in the figure1, figure2 and figure3. More over it is quite unnecessary to use color bars since they just represent one variable.

Tools used in the analysis: Matlab R2010a and Microsoft Excel.

Challenge of the task: Medium.

Time consumption for the task: 12 hours.

Santosh Tirunagari
245577
Santosh.tirunagari@aalto.fi

Impact of other policies and automobiles on caravan policy

Results:

The results were according to the hypothesis made at the beginning of the report, except for the boat policies. That is because of the low status customer (social status D), who use caravans for their mobile homes but not for the interests in the outdoor lifestyles.

Conclusions:

- ✓ Visualization is based on the multivariate data analysis.
- ✓ Figure4 is an example for the high DI-ratio with unnecessary gridlines and color bar. In the other figures the DI-ratio is minimized. Though a single line would be necessary for drawing the bars. But to maintain some elegancy and aesthetic sense, the DI-ratio is optimized.
- ✓ Multiple colors are used in doughnut charts only to discriminate the contributions of each category.
- ✓ Single color is used in the other figures only to show that they belong to same category which obeys the gestalt's laws.
- ✓ 3D graphs are not used in this report. This is because it can generate some lie factors and lead to chart junk and ducks which do not obey the as Tufte's principles.

References:

I. E.R. Tufte. The Visual Display of Quantitative Information (2nd edition).Graphics Press, 2001.
II. http://kdd.ics.uci.edu/databases/tic/tic.html
III. http://en.wikipedia.org/wiki/Chart