PROJECT 1: Data Visualization in Auto Insurance

Seth Howells

seththowells@lewisu.edu

DATA-53000-002, FALL

Data Visualization

Lewis University

1. INTRODUCTION

Predicting *customer value* plays an integral role in any successful marketing model because understanding the *perception* of the product, or object(s), creates opportunities for strategic offerings. *Customer value* refers to the perception of “what a product or service is worth to a customer versus the possible alternatives” [1]. This definition considers the perception of a company’s product or service from the customer’s point of view. While introspective analysis of a company from the customer’s point of view is a wealth of knowledge in itself, exploring customer value from the perspective of profitability makes the most sense for business that hedge against risks, e.g. auto insurance industry.

This paper will approach the same definition [1], but from the company’s profitability point of view, where profitability is the group’s lowest risk and highest premium. Data visualizations created from Tableau serve to gain insight into the customer value perspective.

1. DATA DESCRIPTION

IBM Watson Analytics launched a Kaggle dataset titled “IBM Watson Marketing Customer Value Data,” which included 24 variables in 9,134 instances. TABLE 1 selects the relevant attributes of the dataset for data visualization analysis [2]. A few data processing tasks were completed in Tableau Prep and then connected to Tableau for data visualization, see FIGURES.

Data preprocessing task included removing null values by deletion of the entire row, which was comprised of less than 1% of total records. The other preprocessing task was augmenting the data to create a new Calculated Field as well as convert the “Months since last Claim” data type into a date, see TABLE II.

1. Selected Data Attributes

| Attribute | Type | Example Value | Description |
| --- | --- | --- | --- |
| State | Nominal (string) | Washington | Customer state for auto insurance coverage |
| Customer Lifetime Value | Numeric (real) | 2763.519279 |  |
| Coverage | Ordinal (string) | Basic |  |
| Education | Ordinal (string) | Bachelor | Level of education pursued |
| EmploymentStatus | Nominal (string) | Employed | Current state of employment |
| Gender | Nominal (string) | Female |  |
| Income | Numeric (integer) | 56274 |  |
| Location Code | Nominal (string) | Suburban | Locality type: Urban, suburban, rural |
| Marital Status | Nominal (string) | Married | Single, married, divorced |
| Monthly Premium Auto | Numeric (integer) | 94 | Customer monthly payment |
| Months Since Last Claim | Numeric (integer) | 32 |  |
| Months Since Policy Inception | Numeric (integer) | 5 |  |
| Number of Policies | Numeric (integer) | 8 |  |
| Sales Channel | Nominal (string) | Agent | Agent, Branch, Call Center, Web |
| Total Claim Amount | Numeric (real) | 384.811147 |  |
| Vehicle Class | Nominal (string) | Two-Door Car | Type of vehicle: two/four door, SUV, Truck |
| Vehicle Size | Nominal (string) | Midsize | Size of vehicle |

1. data augmentations

| Attribute | Type | Example Value | Description |
| --- | --- | --- | --- |
| Adjusted Total Claim | Numeric (real) | 398.76 | avg Total Claim – avg Monthly Premium |
| Months since last Claim | Date | 01 | Count of months from last claim date to current claim date |

1. METHODOLOGY

In efforts to make sense of the data, several histograms, conditional charts, and box-plots were created. The goal was to assess the distribution of the data in a fast, yet thorough way. These data visualization techniques confirmed some initial assumptions like: luxury cars are typically more expensive to insure, customer’s income is highly correlated with higher monthly premiums, and that the sample size of higher-education customers are a significant minority among the population. Although confirmation of the initial assumptions could not generate meaningful insight, it gave an overall picture of the composition of the customers, like: suburban residence skews higher than urban and rural – both for different reasons.

Identifying potential outliers was mostly done through box-plots since quartiles are automatically generated. This allowed for quick investigation for instances that were far too far from the population. There were no patterns identified in outliers as the outliers were less than 1 percent of the total sample.

1. FIGURES

Map

Description automatically generated

*FIG 1. Geographical Overview of Total Claim & Monthly Payments*

Chart, bar chart

Description automatically generated

*FIG 2. Customer Lifetime Value cluster and intra-cluster*

Chart

Description automatically generated

*FIG 3. Running months since last Claim*

Chart, box and whisker chart

Description automatically generated

*FIG 4. Total Claim by Education Level*

*Chart, bubble chart

Description automatically generated*

*FIG 5. Avg Monthly Premium: Vehicle Class & Location of Residence*

1. CONCLUSIONS

Business models that depend on hedging risks so that the monthly revenue appropriately exceeds the monthly claim payout. Because the auto insurance industry agrees to set the monthly premium without full knowledge of the customer that is signing the agreement, predicting customer value becomes integral to the process. Seeing the data in new perspective is one way to explore new insight that may play a large role in a future marketing decision.

FIG 1 provides a map of the states the customers reside with a scaling factor. The darker the state, the higher total/average monthly premium/claim. The lighter the states are aggregated by the lowest of premiums and claims. It is important to note that because the standard deviation is within $10.00 per month, a new calculated field was generated which adjusted the average Total Claim. The result of the insight was that Washington and Arizona are the areas with the least monthly premium; however, the adjusted values (left) show only Arizona as the “lighter” states.

FIG 2 attempts to cluster the customers into two groups based on their Lifetime Value. The groups are separated from lower values (cluster 2 – orange) to higher values (cluster 1- blue). The bottom-left figure references the customer residential location type and educational background. It was clear that cluster 2 primarily resides in suburban areas. The bottom-right figure assesses the educational background of cluster 2. It was clear Master and Doctor level customers were far less found than ratios in Urban residence types.

FIG 3 helps describe the likelihood and total amount of a claim, given the customer has driven X amount of months with no claim. The assumption of this figure is that, for every month a customer has driven without submitting a claim, the less likely the customer is to file a new claim with the caveat that the new claim is typically less expensive.

FIG 4 dives deeper into the Education Level with consideration toward Total Claim. The average total claim amount (left) gave a better understanding intra Education Level relationship. Since total claim amount is heavily skewed in relation to Education Level – see initial assumptions in METHODOLOGY- the average total claim gave a unique perspective into the typical claim distribution per educational level achieved. Customers that have achieved Doctor or Masters degree typically have less of a Claim.

FIG 5 illustrates revenue streams, i.e. Monthly Premium, from the perspective of vehicle type and residential location. The size of the objects refer to the monthly premium – bigger the object, the higher the premium. As expected, luxury car and luxury SUV types have the largest premiums; however, it was interesting to see that sports cars and regular SUV are generally much more expensive that a two-door or four-door car in any residential area.

REFERENCES

[1] G. Mahajan, “What is customer value and how can you create it?,” *Journal of Creating Value*, vol. 6, no. 1, pp. 119–121, 2020.

[2] P. Joshi, “IBM Watson Marketing Customer Value Data,” *Kaggle*, 08-May-2019. [Online]. Available: https://www.kaggle.com/pankajjsh06/ibm-watson-marketing-customer-value-data. [Accessed: 01-Nov-2021].