Portland State University

PDXScholar

Engineering and Technology Management Student Projects | Engineering and Technology Management

Winter 2018

# Data Warehousing Class Project Report

Gaya Haciane
*Portland State University*

Chuan Chieh Lu
*Portland State University*

Rassaniya Lerdphayakkarat
*Portland State University*

Rudraxi Mitra
*Portland State University*

Follow this and additional works at: https://pdxscholar.library.pdx.edu/etm_studentprojects

Part of the Business Analytics Commons, and the Databases and Information Systems Commons

Let us know how access to this document benefits you.

## Citation Details

Haciane, Gaya; Lu, Chuan Chieh; Lerdphayakkarat, Rassaniya; and Mitra, Rudraxi, "Data Warehousing Class Project Report" (2018). *Engineering and Technology Management Student Projects*. 1944.
https://pdxscholar.library.pdx.edu/etm_studentprojects/1944

This Project is brought to you for free and open access. It has been accepted for inclusion in Engineering and Technology Management Student Projects by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.

# Assignment #7 Class Project Report

**Course Title:** Data Warehousing
**Course Number:** ETM 538/638
**Instructor:** Mike Freiling, Daniel Sagalowicz
**Term:** Winter 2018
**Student Name:** Gaya Haciane, Chuan Chieh Lu, Rassaniya Lerdphayakkarat, Rudraxi Mitra

ETM OFFICE USE ONLY

Report No.:
Type: Student Project
Note:

## Table of Content

## I- Introduction

Data mining is widely described or defined as the discipline of: "making sense of the data". In today's day and age, the rise of ubiquity of information calls for more advanced and developed techniques to mine the data and come up with insights. Data mining finds applications in many different fields and industries: Whether it is in Embryology, Crops, Elections, or Business Marketing...etc. It is not a wild assumption to consider that every organization in the world has some data mining capabilities or its main activity necessitates it and they have some third party organization doing that for them. One particular area where data mining is really important is in the business world. Being able to find patterns in the data can tell whether the business survives for another couple of years or not. It can make the difference between being a fortune 500 company and bankruptcy and everybody who is interested in growth and sustainability knows that. During the whole course, we learned methodology and did assignments for practicing data mining and data warehousing. In this class project, we try to put to practice as many concepts as those learned in class and apply 3 algorithms from class (1-R, Bayesian, and Instant-based).

## II- The Data

The data set that was used for this project was retrieved from IBM Watson Analytics online community platform where other datasets are made available [1]. This is dataset comes from a car insurance company whose name was undisclosed. The data set has **26 attributes and 9134 records.** It has no missing values and the dependent variable is the attribute: **CLV**, standing for ***customer lifetime value.*** The description of 26 attributes along with their nature (numerical, categorical, answer, question, link) is shown in Appendix A.

**Definition:** Customer lifetime value is a marketing concept that refers to the amount of money that will be made from a customer over its lifetime as a company customer. In its calculation the analyst should be mindful of the **Cost of Customer Acquisition (CAC)**, periodic profit made from this customer over a certain period of time and the duration this customer will still be a customer of the company. **CLV** is popular concept in Banks, insurance companies (cars, health…etc.) and virtually any business.

## III- The Need for the project

### 1. Key Business Objectives

The Key business objectives of this project is to increase the ***Customer Lifetime Value (CLV)*** of customers of a car insurance company. The objective will be met by analyzing the different attributes and how they impact the **CLV.** The project insights will serve in designing predictive analytical methods that will help the business owner tell whether a prospective customer will have a high lifetime value or not and based on that have our client act on some aspects to either keep the CLV high or take action to increase it.

### 2. Key business questions

1. Who are the customers that have the higher customer lifetime value? This can be categorized by (gender, location, age, income, vehicle type, employment...etc).

2. What type of insurance generates the most value by claims?
3. Which vehicles type and size has the most claim amount?
4. What policy type is more profitable?
5. What channel has is the most conversion rate?
6. Who are the customers that have the highest risk of recurring claims? (categorize them by education)
7. What are expiration date of different insurance policies by their coverage type?
8. What are coverage type of insurance that have most complains?
9. What is the number of complains of a certain policy types ?
10. What are the months since last inception and months since last claim for a certain no of policy types?

3. **Concepts the Organization is already using to analyze the data**

This dataset was made available by IBM Watson analytics for, mostly, academic reasons. The name of insurance company as specified earlier was no disclosed. The tool that is used to analyze the data is **IBM Watson Analytics which** is an advanced data analysis and visualization solution in the cloud and the concepts involved are: Natural language dialogue, Automated predictive analytics, One-click analysis, Smart data discovery, Simplified analysis, Accessible advanced analytics, Self-service dashboards.

## IV- Procedure of analysis

1. **Key attributes to use**

In this project the key attributes to use are: **VehicleClass**, Monthly premium amount called **Premium**, and type of insurance coverage called **Coverage**. We use three different algorithms, but all of three key attributes were used in the 3-different algorithm as well.

2. **Any bucketing you plan to use for key attributes**

Two attributes **(Customer Lifetime Value and Premium)** that were used in all the analyses were bucketed. The bucketing happened twice. While running the Bayesian Naive algorithm we made the following buckets:

| **Bucketing#1** | **Bucketing#2** |
|---|---|
| <u>Customer lifetime value (CLV)</u><br>Bucket A: CVL <= $5,000 per year<br>Bucket B: $5000 < CVL <= $20000 per year<br>Bucket C: $20000 < CVL <= $40000 per year<br>Bucket D: $40000 < CVL <= $60000 per year<br>Bucket E: $60000 < CVL per year<br><u>Monthly premium buckets (Premium)</u><br>Low: premium<= $100<br>Medium: $100< premium <=$150<br>High: $150 < premium | <u>Customer lifetime value (CLV)</u><br>Bucket A: CVL <= $3,000 per year<br>Bucket B: $3,000 < CVL <= $6,000 per year<br>Bucket C: $6,000 < CVL <= $12,000 per year<br>Bucket D: $12,000 < CVL <= $24,000 per year<br>Bucket E: $24,000 < CVL per year<br><u>Monthly premium buckets (Premium)</u><br>Low: premium <= $100<br>Medium: $100 < premium <= $150<br>Mid-high: $150 < premium <= $200<br>High: $200 < premium |

The need for bucketing again stems from the fact that the first buckets did not give satisfying answers and therefore needed to be checked out. The results of our analyses that we present here are the ones associated with **Bucketing#2**

3. **Algorithms you think are worth trying. (Only in the class are allowed)**
   Algorithms that are worth trying are: R1, Bayesian Naive, and Instant based classification.

4. **Evaluation criteria**

Depending on the algorithm, evaluation criteria might change, but the universal: Low error rate, high support and high probability should be the main evaluation criteria. Therefore, a good rule will be one that has a lot of support (big enough sample to study it), has low error and its probability of happenstance is considerable high.

## V- Applying the Algorithms

1. **1-R Rule (Bucketing#2)**

After getting the new buckets, we used 1-R to find the best rules to predict CLV based on the three attributes as mentioned. We did 1-R in a single condition, two conditions, and three conditions. For the single condition, we did calculate the error as you can see in Appendix B. The two and three conditions R1, we showed the best rules with the support, and accuracy as following. We used count of CLV buckets instead of average the CLV because CLV has huge range of data which will not provide insight data where the majority is from.

<u>**From the Pivot table**</u>

<u>The best 1-condition rule:</u>

1). if **Premium** = high, then CVL Bucket = D, error =56.27%
2). if **Coverage** = extended , then CVL Bucket = C, error = 55.22%
3). if **VehicleClass** = Luxury Car , then CVL Bucket = D, error = 54.6%
***Note: the errors from 1-condition rule are high because there are five bucket which means it has less percent to have the same result from one condition.***

<u>The best 2-condition rule:</u>

1). if **Coverage** = Premium & **Premium** = high, Then CLV = C
(support = 31, confidence = 31/48, accuracy = 64.6%)
2). if **Coverage** = Premium & **VehicleClass** = Luxury SUV, Then CLV = C
(support = 17 , confidence = 17/26, accuracy = 65.4%)
3). if **Premium** = low & **VehicleClass** = Sports Car, Then CLV = C
(support = 8, confidence = 8/12, accuracy = 66.7%)
***Note: in finding support and accuracy, for each rule, we found from Pivot table by adding sup-row to show counting of each CLV in each condition.***

<u>The best 3-condition rule:</u>

1). if **Coverage** = Premium & **VehicleClass** = Luxury SUV & **SalesChannel** = Agent, Then CLV = C
(support = 16 , confidence = 16/19, accuracy = 84.2%)
2). if **Premium** = low & **Vehicle Class** = Sports car & **EmploymentStatus** = Employed, Then CLV = C (support = 6, confidence = 6/7, accuracy = 85.7%)

3). if **Coverage** = Premium & **Premium** = high & **SalesChannel** = Agent, Then CLV = C (support = 25, confidence = 25/28, accuracy = 89.3%)
***Note: in finding support and accuracy, for each rule, we used the pivot tables form 2-condition and filtered the third condition to find the best rules with high accuracy.***

### 2. Bayesian Naive (Bucketing#2)

The Bayesian model was run to find the value of CLV associated with each combination of values of the attributes **(VehicleClass, Coverage and Premium)** along with returning the probability of accurate decision for each decision.
The full data will be presented in an Excel file that will be attached with this report. Also, it can be found at the Appendix C. Following is an example of one of the best rules that we can come up with by running the Bayes Naive Algorithm.

| Vehicle Class | Coverage Type | Premium | Decision | Probability |
|---|---|---|---|---|
| Luxury Car | Premium | mid-high | D | 85.50% |
| Luxury SUV | Premium | mid-high | D | 83.40% |
| SUV | Premium | mid-high | D | 69.60% |

Once the Bayes model is set up, The insurance company, whenever faced with a new customer profile, they can pick their data and enter them to the model and then the model will be able to predict with relatively good accuracy in what CLV bucket category this customer will be falling and hence will help the insurance company take action based on that.

### 3. Instant based Classification (Bucketing#2)

In the instant-based classification method, the second buckets of the data were used. Only three attributes were considered: **VehicleClass, Coverage type, and Premium Amount.** A few instances (records) of those variables were taken to run the algorithm. As seen in class, the Instant-based classification can turn out to be very time-consuming with long running times when you have large amounts of data. The full data will be presented in an Excel file that will be attached with this report.

The training set is shown in the table below. In interpretation of the results, only 14 out of 72 **(20%)** possible combinations of the data take on one **CLV** value without ambiguity. (Shown across).

It is clear from the results that this Algorithm is not adapted for all possible variables. It appears to do well when **Premium** Coverage value is selected. As the table shows.

| No. | Observation | | | Sequence |
|---|---|---|---|---|
| 1 | Two-Door | high | Premium | D |
| 2 | Four-Door | high | Premium | D |
| 3 | SUV | low | Premium | D |
| 4 | SUV | med-high | Extended | B |
| 5 | SUV | med-high | Premium | B |
| 6 | Luxury Car | low | Premium | D |
| 7 | Luxury Car | medium | Extended | 2E |
| 8 | Luxury Car | medium | Premium | D |
| 9 | Luxury Car | med-high | Extended | 2E |
| 10 | Luxury Car | med-high | Premium | D |
| 11 | Luxury Car | high | Extended | 2E |
| 12 | Luxury Car | high | Premium | D |
| 13 | Sport car | low | Premium | E |
| 14 | Sport car | medium | Premium | E |

This Algorithm despite its ability to work very well with the data takes a long running time and performed poorly, and therefore we do not recommend using it to analyze this data with no automatic system. The recommendations we can infer from the results to make the algorithm more robust as far as analyzing out insurance company data are the following:

1- Experiment with different bucketing schemes.

2- Make the training sample a bit bigger. (which could be very time consuming if done manually).

| Vehicle Class | Premium Amount | Coverage | CLV | 1 | 1 | 1 | DIST |
|---|---|---|---|---|---|---|---|
| Two-Door | low | Extended | A | 1 | 1 | 0 | 2 |
| Four-Door | low | Basic | A | 1 | 1 | 1 | 3 |
| Four-Door | low | Extended | A | 1 | 1 | 0 | 2 |
| Four-Door | low | Basic | A | 1 | 1 | 1 | 3 |
| SUV | mid-high | Premium | B | 1 | 1 | 1 | 3 |
| Four-Door | low | Extended | B | 1 | 1 | 0 | 2 |
| Two-Door | low | Extended | B | 1 | 1 | 0 | 2 |
| Four-Door | low | Extended | B | 1 | 1 | 0 | 2 |
| Two-Door | low | Extended | C | 1 | 1 | 0 | 2 |
| SUV | medium | Basic | C | 1 | 1 | 1 | 3 |
| Two-Door | low | Basic | C | 1 | 1 | 1 | 3 |
| Two-Door | low | Basic | C | 1 | 1 | 1 | 3 |
| SUV | medium | Basic | D | 1 | 1 | 1 | 3 |
| Luxury Car | high | Premium | D | 1 | 0 | 1 | 2 |
| Four-Door | low | Basic | D | 1 | 1 | 1 | 3 |
| Luxury SU\ | high | Extended | D | 1 | 0 | 0 | 1 |
| Luxury Car | high | Extended | E | 1 | 0 | 0 | 1 |
| Luxury Car | high | Extended | E | 1 | 0 | 0 | 1 |
| Luxury SU\ | high | Extended | E | 1 | 0 | 0 | 1 |
| Sports Car | mid-high | Premium | E | 0 | 1 | 1 | 2 |

## VI- Conclusion

In this class project, an insurance company data set was analyzed. The team worked on applying all the important algorithms learned in class, and we tried to put to practice all the different concepts and techniques that were seen. The algorithms performed differently, which puts in perspective the idea of using the right algorithms for the the right application. Insights from this class project are summarized in what follows:

a) **Insights regarding the methods:**

- Algorithms can be application dependent.
- Bucketing can change the results of your analysis and therefore, one has got to be mindful of selecting robust and rational bucketing schemes to ensure the data is not completely skewed.
- Increasing the number of attributes used in an analysis, in most cases (in this project) increases the accuracy of prediction, but one has to be mindful to select just the right number of attributes. Overfitting issues might rise, and that will make the analysis insights basically useless.

**b) Insights regarding the results of our application**

- Depending on the application, our client can use any algorithm to predict the CLV of prospective customers.
- Ex: 1-R 3-condition can be used to target new customers offering premium coverage, with high monthly premium amount and reach out to them via agent will lead to C-level CLV.
- The algorithms' results can either be used by the insurance company to either improve their **Customer Relationship Management**, or even to acquire new customers.
- Once the models are set up, our client can use them to answer any of the business questions they might have.
- The attributes that our client should focus on should be: VehicleClass, Coverage, Premium amount, and Sales Channel.

**VII- References**

[1] "SAMPLE DATA: Marketing Customer Value Analysis," *IBM Analytics Communities*, 11-Apr-2015. [Online]. Available: https://www.ibm.com/communities/analytics/watson-analytics-blog/marketing-customer-value-analysis/. [Accessed: 09-Mar-2018].

[2] "IBM Watson Analytics," *IBM Watson Analytics - Overview - United States*, 10-Mar-2018. [Online]. Available: https://www.ibm.com/us-en/marketplace/watson-analytics. [Accessed: 09-Mar-2018].

[3] Witten, I., Frank, Eibe, & Hall, Mark A. (2011). Data mining : Practical machine learning tools and techniques (3rd ed., Morgan Kaufmann series in data management systems). Burlington, MA: Morgan Kaufmann.

## VIII- Appendix

### Appendix A: The description of 26 attributes

The attributes along with their nature are shown in the following table:

| Attribute | Description | Type | Nature |
|---|---|---|---|
| **Customer** | Different customers with their own ID | Text and Integer | Link |
| **State** | Name of states in which insurance is sold | Text | Answer |
| **Customer Lifetime Value (CLV)** | The time period since a particular person has been paying premiums | Currency | Key-Answer |
| **Response** | No or yes response to the coverage of insurance type | Text | Answer |
| **Coverage** | The coverage type of insurance | Text | Answer |
| **Education** | The education of customers buying the insurance | Text | Answer |
| **Effective to Date** | The time period until the insurance is active | Date | Answer |
| **Employment Status** | The employment status of customer | Text | Answer |
| **Gender** | The gender of each customer buying insurance | Text | Answer |
| **Income** | The income of customers buying insurance | Currency | Answer |
| **Location Code** | The location of each customer | Text | Answer |
| **Marital Status** | The marital status of each customer | Text | Answer |
| **Monthly Premium Auto** | The insurance premiums paid for each auto | Integer | Answer |
| **Premium** | The amount paid for an insurance policy | Text | Answer |
| **Months Since Last Claim** | The number of months passed since the insurance is claimed. | Integer | Answer |

| | | | |
|---|---|---|---|
| **Months Since Policy Inception** | The insurance was first purchased | Integer | Answer |
| **Number of Open Complaints** | The number of complaints by each customer | Integer | Answer |
| **Number of Policies** | The number of policies sold by each customer | Integer | Answer |
| **Policy Type** | The types of insurance policy | Text | Answer |
| **Policy** | Name of policy | Text | Answer |
| **Renew Offer Type** | The type of offer | Text | Answer |
| **Sales Channel** | The channel through which insurance is sold | Text | Answer |
| **Total Claim Amount** | Claimed amount of each policy type of insurance | Currency | Answer |
| **Vehicle Class** | The class of vehicles being most claimed | Text | Answer |
| **Vehicle Size** | The size of vehicles that has auto insurance | Text | Answer |

## Appendix B: Data and Pivot tables of R1

### The training Data:

| Il | Custo | State | Customer | CL | Respo | Covera | Education | Effective | Employment | G | Incon | Incon | Location | Marital : | Montl | premiu | Months Since La | Months Since Policy | Number of Open Co | Number of F | Policy Ty | Policy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | FQ61281 | Oregon | $ 83,325.38 | E | No | Extended | High School or Belc | 2011/1/31 | Employed | M | 58958 | d | Suburban | Married | 231 | high | 31 | 74 | 0 | 2 | Personal Auto | Persona |
| 2 | YC54142 | Washington | $ 74,228.52 | E | No | Extended | High School or Belc | 2011/1/26 | Unemployed | M | 0 | a | Suburban | Single | 242 | high | 1 | 34 | 0 | 2 | Personal Auto | Persona |
| 3 | BP23267 | California | $ 73,225.96 | E | No | Extended | Bachelor | 2011/2/9 | Employed | F | 39547 | c | Suburban | Married | 202 | high | 11 | 21 | 0 | 2 | Personal Auto | Persona |
| 4 | KH55886 | Oregon | $ 67,907.27 | E | No | Premium | Bachelor | 2011/2/5 | Employed | M | 78310 | d | Rural | Married | 192 | mid-high | 34 | 18 | 1 | 2 | Personal Auto | Persona |
| 5 | SK66747 | Washington | $ 66,025.75 | E | No | Basic | Bachelor | 2011/2/22 | Employed | M | 33481 | c | Suburban | Single | 188 | mid-high | 28 | 46 | 0 | 2 | Personal Auto | Persona |
| 6 | FB95288 | California | $ 64,618.76 | E | No | Extended | High School or Belc | 2011/1/17 | Unemployed | M | 0 | a | Suburban | Married | 217 | high | 14 | 40 | 1 | 2 | Personal Auto | Persona |
| 7 | AZ84403 | Oregon | $ 61,850.19 | E | No | Extended | College | 2011/2/4 | Unemployed | F | 0 | a | Suburban | Married | 238 | high | 19 | 29 | 0 | 2 | Personal Auto | Persona |
| 8 | US30122 | California | $ 61,134.68 | E | No | Basic | College | 2011/2/28 | Unemployed | M | 0 | a | Suburban | Single | 198 | mid-high | 2 | 75 | 0 | 2 | Corporate Autc | Corpor |
| 9 | JT47995 | Arizona | $ 60,556.19 | E | No | Extended | College | 2011/1/1 | Unemployed | F | 0 | a | Suburban | Married | 204 | high | 35 | 45 | 0 | 2 | Personal Auto | Persona |
| 10 | EN65835 | Arizona | $ 58,753.88 | E | No | Premium | Bachelor | 2011/1/6 | Employed | F | 24964 | b | Suburban | Married | 185 | mid-high | 0 | 84 | 0 | 2 | Personal Auto | Persona |
| 11 | XF89906 | Arizona | $ 58,207.13 | E | No | Extended | High School or Belc | 2011/1/13 | Disabled | M | 29295 | c | Suburban | Married | 219 | high | 25 | 50 | 0 | 2 | Personal Auto | Persona |
| 12 | OM82309 | California | $ 58,166.55 | E | No | Basic | Bachelor | 2011/2/27 | Employed | M | 61321 | d | Rural | Single | 186 | mid-high | 0 | 30 | 1 | 2 | Personal Auto | Persona |
| 13 | JZ23377 | Oregon | $ 57,520.50 | E | No | Premium | College | 2011/1/20 | Employed | F | 48367 | c | Suburban | Married | 161 | mid-high | 10 | 34 | 0 | 2 | Personal Auto | Persona |
| 14 | DU50092 | Oregon | $ 56,675.94 | E | No | Premium | College | 2011/1/24 | Employed | F | 77237 | d | Suburban | Married | 283 | high | 33 | 93 | 0 | 2 | Personal Auto | Persona |
| 15 | OY68395 | Oregon | $ 55,277.45 | E | No | Basic | High School or Belc | 2011/1/30 | Employed | F | 40740 | c | Suburban | Single | 198 | mid-high | 19 | 60 | 0 | 2 | Personal Auto | Persona |
| 16 | CL79250 | Nevada | $ 52,811.49 | E | No | Basic | Bachelor | 2011/1/8 | Unemployed | M | 0 | a | Suburban | Married | 182 | mid-high | 8 | 70 | 0 | 2 | Corporate Autc | Corpor |
| 17 | AH58807 | Arizona | $ 51,426.25 | E | No | Basic | College | 2011/1/9 | Employed | F | 84650 | d | Urban | Married | 185 | mid-high | 13 | 39 | 3 | 2 | Personal Auto | Persona |
| 18 | KI58952 | California | $ 51,337.91 | E | No | Premium | College | 2011/2/24 | Employed | F | 72794 | d | Rural | Single | 164 | mid-high | 3 | 47 | 1 | 2 | Personal Auto | Persona |
| 19 | LW64678 | California | $ 51,016.07 | E | No | Premium | Master | 2011/2/19 | Employed | F | 25167 | c | Urban | Married | 140 | medium | 3 | 76 | 0 | 2 | Personal Auto | Persona |
| 20 | QT84069 | Oregon | $ 50,568.26 | E | No | Extended | Master | 2011/2/28 | Employed | M | 82081 | d | Urban | Married | 249 | high | 1 | 62 | 0 | 2 | Personal Auto | Persona |
| 21 | BR50492 | Arizona | $ 49,423.80 | E | No | Extended | Bachelor | 2011/1/4 | Employed | M | 85058 | d | Urban | Married | 137 | medium | 34 | 82 | 0 | 2 | Personal Auto | Persona |
| 22 | RP30093 | Oregon | $ 49,221.43 | E | No | Premium | Bachelor | 2011/1/23 | Employed | F | 63035 | d | Suburban | Married | 153 | mid-high | 20 | 97 | 0 | 2 | Personal Auto | Persona |
| 23 | LU42720 | Nevada | $ 48,356.96 | E | No | Extended | College | 2011/2/20 | Employed | M | 52499 | d | Suburban | Divorced | 138 | medium | 0 | 61 | 0 | 2 | Personal Auto | Persona |
| 24 | MJ77630 | Oregon | $ 47,155.63 | E | No | Extended | High School or Belc | 2011/2/10 | Employed | M | 39891 | c | Urban | Married | 133 | medium | 12 | 31 | 0 | 2 | Personal Auto | Persona |
| 25 | CP92616 | Nevada | $ 46,805.22 | E | No | Extended | High School or Belc | 2011/2/25 | Employed | M | 83006 | d | Urban | Married | 235 | high | 8 | 61 | 1 | 2 | Personal Auto | Persona |
| 26 | KB44286 | Oregon | $ 46,770.95 | E | No | Basic | High School or Belc | 2011/2/1 | Employed | F | 64403 | d | Rural | Single | 198 | mid-high | 11 | 86 | 0 | 2 | Corporate Autc | Corpor |
| 27 | DM76654 | California | $ 46,611.87 | E | No | Extended | College | 2011/2/6 | Employed | M | 22022 | b | Suburban | Single | 136 | medium | 35 | 1 | 0 | 2 | Personal Auto | Persona |
| 28 | GV41938 | California | $ 46,302.08 | E | No | Premium | High School or Belc | 2011/2/6 | Unemployed | F | 0 | a | Suburban | Married | 151 | mid-high | 33 | 94 | 1 | 2 | Personal Auto | Persona |
| 29 | OK56965 | California | $ 45,708.65 | E | No | Basic | Bachelor | 2011/1/19 | Employed | F | 31264 | c | Urban | Divorced | 198 | mid-high | 9 | 54 | 1 | 2 | Personal Auto | Persona |
| 30 | ZF84966 | Nevada | $ 44,856.11 | E | No | Extended | Doctor | 2011/2/22 | Employed | F | 61675 | d | Urban | Married | 123 | medium | 35 | 54 | 0 | 2 | Corporate Autc | Corpor |
| 31 | CP85232 | Arizona | $ 44,795.47 | E | No | Extended | College | 2011/2/4 | Employed | M | 58778 | d | Rural | Married | 126 | medium | 6 | 62 | 0 | 2 | Special Auto | Special |
| 32 | AB31813 | Washington | $ 44,771.30 | E | No | Extended | High School or Belc | 2011/2/12 | Unemployed | M | 0 | a | Suburban | Married | 131 | medium | 20 | 59 | 0 | 2 | Personal Auto | Persona |
| 33 | SD41771 | California | $ 44,520.14 | E | No | Premium | High School or Belc | 2011/1/26 | Employed | M | 49259 | c | Suburban | Single | 144 | medium | 35 | 36 | 0 | 2 | Personal Auto | Persona |
| 34 | XZ62712 | Washington | $ 44,468.02 | E | No | Extended | High School or Belc | 2011/1/29 | Employed | F | 32948 | c | Suburban | Single | 127 | medium | 20 | 46 | 1 | 2 | Personal Auto | Persona |

**The Pivot Table for Premium**

| Total | Max count | The rest | Decision | Error |
|---|---|---|---|---|
| 510 | 223 | 287 | D | 0.562745 |
| 5990 | 2311 | 3679 | B | 0.61419 |
| 2634 | 1082 | 1552 | C | 0.589218 |
| 9134 | 3616 | 5518 | Average | 0.604116 |

**The Pivot Table for Coverage**

| Total | Max count | The rest | Decision | Error |
|---|---|---|---|---|
| 5568 | 2045 | 3523 | B | 0.632723 |
| 2742 | 1228 | 1514 | C | 0.552152 |
| 824 | 340 | 484 | C | 0.587379 |
| 9134 | 3613 | 5521 | Average | 0.604445 |

**The Pivot Table for Vehicle Class**

| Total | Max count | The rest | Decision | Error |
|---|---|---|---|---|
| 4621 | 1783 | 2838 | B | 0.614153 |
| 163 | 74 | 89 | D | 0.546012 |
| 184 | 79 | 105 | D | 0.570652 |
| 484 | 219 | 265 | C | 0.547521 |
| 1796 | 706 | 1090 | C | 0.606904 |
| 1886 | 703 | 1183 | B | 0.627253 |
| 9134 | 3564 | 5570 | Average | 0.60981 |

## Appendix C : Bayesian Model Probabilities Data
**Vehicle Class:**

| Label | Coverage type | CLV | Count | CLVCount | Prob |
|---|---|---|---|---|---|
| Four-Door-Car\|A | Four-Door-Car | A | 1060 | 1506 | 0.70385 |
| Four-Door-Car\|B | Four-Door-Car | B | 1783 | 3248 | 0.54895 |
| Four-Door-Car\|C | Four-Door-Car | C | 1333 | 2929 | 0.4551 |
| Four-Door-Car\|D | Four-Door-Car | D | 351 | 1093 | 0.32113 |
| Four-Door-Car\|E | Four-Door-Car | E | 94 | 358 | 0.26257 |
| Luxury Car\|A | Luxury Car | A | 0 | 1506 | 0 |
| Luxury Car\|B | Luxury Car | B | 1 | 3248 | 0.00031 |
| Luxury Car\|C | Luxury Car | C | 62 | 2929 | 0.02117 |
| Luxury Car\|D | Luxury Car | D | 74 | 1093 | 0.0677 |
| Luxury Car\|E | Luxury Car | E | 26 | 358 | 0.07263 |
| Luxury SUV\|A | Luxury SUV | A | 0 | 1506 | 0 |
| Luxury SUV\|B | Luxury SUV | B | 0 | 3248 | 0 |
| Luxury SUV\|C | Luxury SUV | C | 78 | 2929 | 0.02663 |
| Luxury SUV\|D | Luxury SUV | D | 79 | 1093 | 0.07228 |
| Luxury SUV\|E | Luxury SUV | E | 27 | 358 | 0.07542 |
| Sport Car\|A | Sport Car | A | 0 | 1506 | 0 |
| Sport Car\|B | Sport Car | B | 155 | 3248 | 0.04772 |
| Sport Car\|C | Sport Car | C | 219 | 2929 | 0.07477 |
| Sport Car\|D | Sport Car | D | 64 | 1093 | 0.05855 |
| Sport Car\|E | Sport Car | E | 46 | 358 | 0.12849 |
| SUV\|A | SUV | A | 1 | 1506 | 0.00066 |
| SUV\|B | SUV | B | 606 | 3248 | 0.18658 |
| SUV\|C | SUV | C | 706 | 2929 | 0.24104 |
| SUV\|D | SUV | D | 353 | 1093 | 0.32296 |
| SUV\|E | SUV | E | 130 | 358 | 0.36313 |
| Two Door-Car\|A | Two Door-Car | A | 445 | 1506 | 0.29548 |
| Two Door-Car\|B | Two Door-Car | B | 703 | 3248 | 0.21644 |
| Two Door-Car\|C | Two Door-Car | C | 531 | 2929 | 0.18129 |
| Two Door-Car\|D | Two Door-Car | D | 172 | 1093 | 0.15737 |
| Two Door-Car\|E | Two Door-Car | E | 35 | 358 | 0.09777 |

## Coverage Type:

| Label | Coverage type | CLV | Count | CLVCount | Prob |
|---|---|---|---|---|---|
| Basic\|A | Basic | A | 1403 | 1506 | 0.93161 |
| Basic\|B | Basic | B | 2045 | 3248 | 0.62962 |
| Basic\|C | Basic | C | 1361 | 2929 | 0.46466 |
| Basic\|D | Basic | D | 593 | 1093 | 0.54254 |
| Basic\|E | Basic | E | 166 | 358 | 0.46369 |
| Extended\|A | Extended | A | 103 | 1506 | 0.06839 |
| Extended\|B | Extended | B | 971 | 3248 | 0.29895 |
| Extended\|C | Extended | C | 1228 | 2929 | 0.41926 |
| Extended\|D | Extended | D | 299 | 1093 | 0.27356 |
| Extended\|E | Extended | E | 141 | 358 | 0.39385 |
| Premium\|A | Premium | A | 0 | 1506 | 0 |
| Premium\|B | Premium | B | 232 | 3248 | 0.07143 |
| Premium\|C | Premium | C | 340 | 2929 | 0.11608 |
| Premium\|D | Premium | D | 201 | 1093 | 0.1839 |
| Premium\|E | Premium | E | 51 | 358 | 0.14246 |

## Monthly Premium:

| Label | Premium | CLV bucket | Count | CLVCount | Prob |
|---|---|---|---|---|---|
| low\|A | low | A | 1505 | 1506 | 0.999335989 |
| low\|B | low | B | 2311 | 3248 | 0.711514778 |
| low\|C | low | C | 1649 | 2929 | 0.562990782 |
| low\|D | low | D | 418 | 1093 | 0.382433669 |
| low\|E | low | E | 107 | 358 | 0.298882682 |
| medium\|A | medium | A | 1 | 1506 | 0.000664011 |
| medium\|B | medium | B | 913 | 3248 | 0.281096059 |
| medium\|C | medium | C | 1082 | 2929 | 0.369409355 |
| medium\|D | medium | D | 452 | 1093 | 0.413540714 |
| medium\|E | medium | E | 186 | 358 | 0.519553073 |
| mid-high\|A | mid-high | A | 0 | 1506 | 0 |
| mid-high\|B | mid-high | B | 24 | 3248 | 0.007389163 |
| mid-high\|C | mid-high | C | 137 | 2929 | 0.046773643 |
| mid-high\|D | mid-high | D | 160 | 1093 | 0.146386093 |
| mid-high\|E | mid-high | E | 28 | 358 | 0.078212291 |
| top \| A | top | A | 0 | 1506 | 0 |
| top \| B | top | B | 0 | 3248 | 0 |
| top \| C | top | C | 61 | 2929 | 0.020826221 |
| top \| D | top | D | 63 | 1093 | 0.057639524 |
| top \| E | top | E | 37 | 358 | 0.103351955 |

## CLV Bucket:

| CLVBucket | CLVCount | Total | Probability |
|---|---|---|---|
| A | 1506 | 9134 | 0.164878476 |
| B | 3248 | 9134 | 0.355594482 |
| C | 2929 | 9134 | 0.320670024 |
| D | 1093 | 9134 | 0.119662798 |
| E | 358 | 9134 | 0.039194219 |

## Predictive Model:

| | Class | Coverage | Premium | CLV Bucket | Product (x $10^{-8}$) | Likelihood (%) |
|---|---|---|---|---|---|---|
| Observation: | Luxury Car | Premium | mid-high | | | |
| | | | | | | |
| A | 0 | 0 | 0 | 0.164878476 | 0 | 0.0% |
| B | 0.000307882 | 0.071428571 | 0.007389163 | 0.355594482 | 5.778381058 | 0.0% |
| C | 0.021167634 | 0.116080574 | 0.046773643 | 0.320670024 | 3685.457633 | 14.5% |
| D | 0.067703568 | 0.18389753 | 0.146386093 | 0.119662798 | 21809.53613 | 85.5% |
| E | 0.072625698 | 0.142458101 | 0.078212291 | 0.039194219 | 3171.571414 | 12.4% |
| | | | | | 25500.77215 | 100.0% |

## Appendix D: Full Table for Instant Based Learning

| | Observation | | | Sequence | Decision | Error |
|---|---|---|---|---|---|---|
| 1 | Two-Door | low | Extended | A-A-B-B-B-C | B | 50% |
| 2 | Two-Door | low | Basic | A-A-C-C | A-C | 50% |
| 3 | Two-Door | low | Premium | 4A-4B-3C-D | A-B | 67% |
| 4 | Two-Door | medium | Extended | 2A-3B-2C-D | B | 62.50% |
| 5 | Two-Door | medium | Basic | C-D | C-D | 50% |
| 6 | Two-Door | medium | Premium | B-C-2D | D | 50% |
| 7 | Two-Door | med-high | Extended | 2A-3B-1C | B | 50% |
| 8 | Two-Door | med-high | Basic | 2A-3C-1D | C | 50% |
| 9 | Two-Door | med-high | Premium | B-D | B-D | 50% |
| 10 | Two-Door | high | Extended | 2A-3B-C-D | B | 58% |
| 11 | Two-Door | high | Basic | 2A-3C-2D | C | 58% |
| 12 | Two-Door | high | Premium | D | D | 0% |
| 13 | Four-Door | low | Extended | 2A-3B-C | B | 50% |
| 14 | Four-Door | low | Basic | 2A-2C | A-C | 50% |
| 15 | Four-Door | low | Premium | 4A-4B-3C-D | A-B | 67% |
| 16 | Four-Door | medium | Extended | 2A-3B-2C-D | B | 62.50% |
| 17 | Four-Door | medium | Basic | C-D | C-D | 50% |
| 18 | Four-Door | medium | Premium | B-C-2D | D | 50% |
| 19 | Four-Door | med-high | Extended | 2A-3B-1C | B | 50% |
| 20 | Four-Door | med-high | Basic | 2A-3C-D | C | 50% |
| 21 | Four-Door | med-high | Premium | B-D | B-D | 50% |
| 22 | Four-Door | high | Extended | 2A-3B-C-D | B | 57% |
| 23 | Four-Door | high | Basic | 2A-3C-2D | C | 57% |
| 24 | Four-Door | high | Premium | D | D | 0% |
| 25 | SUV | low | Extended | 2A-3B-C | B | 50% |
| 26 | SUV | low | Basic | 2A-3C-D | C | 50% |
| 27 | SUV | low | Premium | D | D | 0% |
| 28 | SUV | medium | Extended | C-D | C-D | 50% |
| 29 | SUV | medium | Basic | C-D | C-D | 50% |
| 30 | SUV | medium | Premium | B-C-D | B-C-D | 66% |
| 31 | SUV | med-high | Extended | B | B | 0% |
| 32 | SUV | med-high | Basic | B-C-D | B-C-D | 66% |
| 33 | SUV | med-high | Premium | B | B | 0% |
| 34 | SUV | high | Extended | 2A-4B-2C-2D | B | 60% |
| 35 | SUV | high | Basic | C-D | C-D | 50% |
| 36 | SUV | high | Premium | B-D | B-D | 50% |
| 37 | Luxury Car | low | Extended | 2A-3B-C-2E | B | 54% |
| 38 | Luxury Car | low | Basic | 2A-2C-D | A-C | 60% |
| 39 | Luxury Car | low | Premium | D | D | 0% |
| 40 | Luxury Car | medium | Extended | 2E | E | 0% |
| 41 | Luxury Car | medium | Basic | C-D | C-D | 50% |

| | | | | | | |
|---|---|---|---|---|---|---|
| 42 | Luxury Car | medium | Premium | D | D | 0% |
| 43 | Luxury Car | med-high | Extended | 2E | E | 0% |
| 44 | Luxury Car | med-high | Basic | 2A-3C-3D-2E | C-D | 70% |
| 45 | Luxury Car | med-high | Premium | D | D | 0% |
| 46 | Luxury Car | high | Extended | 2E | E | 0% |
| 47 | Luxury Car | high | Basic | 2E-D | E | 30% |
| 48 | Luxury Car | high | Premium | D | D | 0% |
| 49 | LuxurySUV | low | Extended | 2A-3B-C-D-E | B | 62% |
| 50 | LuxurySUV | low | Basic | 4A-3B-3C-2D-E | A | 70% |
| 51 | LuxurySUV | low | Premium | 4A-4B-3C-3D-2E | A-B | 75% |
| 52 | LuxurySUV | medium | Extended | D-E | D-E | 50% |
| 53 | LuxurySUV | medium | Basic | C-D | C-D | 50% |
| 54 | LuxurySUV | medium | Premium | B-C-3D-2E | D | 57% |
| 55 | LuxurySUV | med-high | Extended | D-E | D-E | 50% |
| 56 | LuxurySUV | med-high | Basic | 2A-3C-3D-E | C-D | 67% |
| 57 | LuxurySUV | med-high | Premium | B-2D-2E | D-E | 60% |
| 58 | LuxurySUV | high | Extended | D-E | D-E | 50% |
| 59 | LuxurySUV | high | Basic | D-E | D-E | 50% |
| 60 | LuxurySUV | high | Premium | 2D-E | D | 33% |
| 61 | Sport car | low | Extended | 2A-3B-C | B | 50% |
| 62 | Sport car | low | Basic | 2A-2C-D | A-C | 60% |
| 63 | Sport car | low | Premium | E | E | 0% |
| 64 | Sport car | medium | Extended | C-D-E | C-D-E | 33% |
| 65 | Sport car | medium | Basic | C-D | C-D | 50% |
| 66 | Sport car | medium | Premium | E | E | 0% |
| 67 | Sport car | med-high | Extended | 2A-3B-C-D-4E | E | 63% |
| 68 | Sport car | med-high | Basic | 2A-3C-2D-E | C | 62.50% |
| 69 | Sport car | med-high | Premium | D-3E | E | 25% |
| 70 | Sport car | high | Extended | 2A-3B-C-D | B | 57% |
| 71 | Sport car | high | Basic | 2A-3C-4D-4E | D-E | 70% |
| 72 | Sport car | high | Premium | D-E | D-E | 50% |