Homework 4

Group 1

Contents

Prepared for:

Dr. Nathan Bastian

City University of New York, School of Professional Studies - Data 621

Prepared by:

Group 1

Senthil Dhanapal

Yadu Chittampalli

Christophe Hunt

1 Introduction

Consumers who own a car are often required to purchase car insurance to protect themselves from serious financial repercussions of being involved in a car accident. Insurance Providers must determine the risk of the offering insurance converage to a new customer through accurate statistical models that evaluate the risk. Since Insurance Providers are motivated by collecting the maximum amount of revenue from consumers while returning the lowest amount in accident claims, the statistical modeling provides Insurance Providers with insight into the consumers behavior and the most appropriate pricing schemes[1].

2 Statement of the Problem

The purpose of this report is to develop statistical models to make inference into the likelihood of a customer being involved in a car accident and the cost associated of a customer being involved in a car accident.

3 Data Exploration

3.1 Variables Explained

The variables provided in our evaluation data set our explained below:

Variable Code	Definition
INDEX	Identification Variable (do not use)
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO
TARGET_AMT	If car was in a crash, what was the cost
AGE	Age of Driver
BLUEBOOK	Value of Vehicle
CAR_AGE	Vehicle Age
CAR_TYPE	Type of Car
CAR_USE	Vehicle Use
CLM_FREQ	# Claims (Past 5 Years)
EDUCATION	Max Education Level
HOMEKIDS	# Children at Home
HOME_VAL	Home Value
INCOME	Income
KIDSDRIV	# Driving Children
MSTATUS	Marital Status
MVR_PTS	Motor Vehicle Record Points
OLDCLAIM	Total Claims (Past 5 Years)
PARENT1	Single Parent
RED_CAR	A Red Car
REVOKED	License Revoked (Past 7 Years)
SEX	Gender
TIF	Time in Force
TRAVTIME	Distance to Work
URBANICITY	Home/Work Area
YOJ	Years on Job

[1] "Insider Information: How Insurance Companies Measure Risk - Insurance Companies.com." Insurance Companiescom. N.p., n.d. Web. 06 Nov. 2016.

Histograms of most of our variables have been plotted below so that distribution can be visualized.

Table 1 : Descriptive Statistics
24 Variables 8161 Observations

TARGET_AMT

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
8161	0	1949	0.601	1504	2574	0	0	0	0	1036	4904	6452

```
lowest :      0.00000     30.27728     58.53106     95.56732    108.74150
highest:  73783.46592  77907.43028  78874.19056  85523.65335 107586.13616
```

KIDSDRIV

n	missing	distinct	Info	Mean	Gmd
8161	0	5	0.318	0.2	0.3

```
lowest : 0 1 2 3 4, highest: 0 1 2 3 4
```

```
0 (7180, 0.880), 1 (636, 0.078), 2 (279, 0.034), 3 (62, 0.008), 4 (4, 0.000)
```

AGE

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
8155	6	60	0.999	45	10	30	34	39	45	51	56	59

```
lowest : 16 17 18 19 20, highest: 72 73 76 80 81
```

HOMEKIDS

n	missing	distinct	Info	Mean	Gmd
8161	0	6	0.723	0.7	1

```
lowest : 0 1 2 3 4, highest: 1 2 3 4 5
```

```
0 (5289, 0.648), 1 (902, 0.111), 2 (1118, 0.137), 3 (674, 0.083), 4 (164, 0.020), 5 (14,
0.002)
```

YOJ

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
7707	454	21	0.989	10	4	0	5	9	11	13	15	15

```
lowest :  0  1  2  3  4, highest: 16 17 18 19 23
```

INCOME

n	missing	distinct
8161	0	6613

```
lowest :         $0      $1,007  $1,022  $1,102
highest: $99,948 $99,959 $99,963 $99,985 $997
```

PARENT1

n	missing	distinct
8161	0	2

```
No (7084, 0.868), Yes (1077, 0.132)
```

HOME_VAL

n	missing	distinct
8161	0	5107

```
lowest :          $0       $100,093 $100,123 $100,226
highest: $99,767  $99,798  $99,815  $99,839  $99,968
```

MSTATUS

n	missing	distinct
8161	0	2

```
Yes (4894, 0.6), z_No (3267, 0.4)
```

SEX

n	missing	distinct
8161	0	2

```
M (3786, 0.464), z_F (4375, 0.536)
```

EDUCATION

n	missing	distinct
8161	0	5

```
lowest : <High School  Bachelors     Masters       PhD           z_High School
highest: <High School  Bachelors     Masters       PhD           z_High School
```

<High School (1203, 0.147), Bachelors (2242, 0.275), Masters (1658, 0.203), PhD (728, 0.089), z_High School (2330, 0.286)

JOB

n	missing	distinct
8161	0	9

```
lowest :               Clerical      Doctor        Home Maker    Lawyer
highest: Lawyer        Manager       Professional  Student       z_Blue Collar
```

Value	Clerical	Doctor	Home Maker	Lawyer	Manager
Frequency	1271	246	641	835	988
Proportion	0.156	0.030	0.079	0.102	0.121

Value	Professional	Student	z_Blue Collar
Frequency	1117	712	1825
Proportion	0.137	0.087	0.224

TRAVTIME

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
8161	0	97	1	33	18	7	13	22	33	44	54	60

```
lowest :   5   6   7   8   9, highest: 103 113 124 134 142
```

CAR_USE

n	missing	distinct
8161	0	2

Commercial (3029, 0.371), Private (5132, 0.629)

BLUEBOOK

n	missing	distinct
8161	0	2789

```
lowest : $1,500 $1,520 $1,530 $1,540 $1,590, highest: $9,950 $9,960 $9,970 $9,980 $9,990
```

TIF

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
8161	0	23	0.961	5	5	1	1	1	4	7	11	13

```
lowest :  1  2  3  4  5, highest: 19 20 21 22 25
```

CAR_TYPE

n	missing	distinct
8161	0	6

```
lowest : Minivan     Panel Truck Pickup      Sports Car  Van
highest: Panel Truck Pickup      Sports Car  Van         z_SUV
```

Minivan (2145, 0.263), Panel Truck (676, 0.083), Pickup (1389, 0.170), Sports Car (907, 0.111), Van (750, 0.092), z_SUV (2294, 0.281)

RED_CAR

n	missing	distinct
8161	0	2

no (5783, 0.709), yes (2378, 0.291)

OLDCLAIM

n	missing	distinct
8161	0	2857

```
lowest : $0     $1,000 $1,008 $1,011 $1,012, highest: $988   $990   $995   $996   $999
```

CLM_FREQ

n	missing	distinct	Info	Mean	Gmd
8161	0	6	0.763	0.8	1

```
lowest : 0 1 2 3 4, highest: 1 2 3 4 5

0 (5009, 0.614), 1 (997, 0.122), 2 (1171, 0.143), 3 (776, 0.095), 4 (190, 0.023), 5 (18,
0.002)
```

REVOKED

n	missing	distinct
8161	0	2

```
No (7161, 0.877), Yes (1000, 0.123)
```

MVR_PTS

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
8161	0	13	0.9	2	2	0	0	0	1	3	5	6

```
lowest :  0  1  2  3  4, highest:  8  9 10 11 13

Value          0     1     2     3     4     5     6     7     8     9    10    11    13
Frequency   3712  1157   948   758   599   399   266   167    84    45    13    11     2
Proportion 0.455 0.142 0.116 0.093 0.073 0.049 0.033 0.020 0.010 0.006 0.002 0.001 0.000
```

CAR_AGE

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
7651	510	30	0.982	8	6	1	1	1	8	12	16	18

```
lowest : -3  0  1  2  3, highest: 24 25 26 27 28
```

URBANICITY

n	missing	distinct
8161	0	2

```
Highly Urban/ Urban (6492, 0.795), z_Highly Rural/ Rural (1669, 0.205)
```

3.2 Exploration of Variables