# Homework 4

*Group 1*

## Contents


**1 Introduction** **2**

**2 Statement of the Problem** **2**

**3 Data Exploration** **2**

3.1 Variables Explained . . . . . 2

3.2 Imputting Missing Values . . . . . 5

3.3 Exploration of Variables . . . . . 7



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 offering insurance coverage to a new customer through accurate statistical models that evaluate the consumers propensity for accidents. Since Insurance Providers are motivated by collecting the maximum amount of revenue from consumers while returning the lowest amount in accident claims, 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 the `Insurance Training Data Set` are 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.

### 3.1.1 Nominal Variables

We first look at our nominal variables and their applicable proportions. Interestingly, we see that in this data set only a quarter of the customer records indicate an accident occurred. Also, the majority of consumers in this data set have no kids at home, are married, more than a high school education but less than a PhD, use their car for private purposes, typically own a SUV or minivan, and also live in an urban environment. This provides an interesting insight to the type of customer this data set represents and should be considered when further interpreting our statistical model. Additionally, we should be mindful of any selection biases in this data set as consumers with extremely risky histories are likely to have not been extended insurance coverage.

Table 2: Table of nominal variables

| **Variable** | **Levels** | $\mathbf{n}$ | $\%$ | $\sum\%$ |
|---|---|---|---|---|
| TARGET_FLAG | 0 | 6008 | 73.6 | 73.6 |
| | 1 | 2153 | 26.4 | 100.0 |
| | all | 8161 | 100.0 | |
| KIDSDRIV | 0 | 7180 | 88.0 | 88.0 |
| | 1 | 636 | 7.8 | 95.8 |
| | 2 | 279 | 3.4 | 99.2 |
| | 3 | 62 | 0.8 | 100.0 |
| | 4 | 4 | 0.0 | 100.0 |
| | all | 8161 | 100.0 | |
| HOMEKIDS | 0 | 5289 | 64.8 | 64.8 |
| | 1 | 902 | 11.1 | 75.9 |
| | 2 | 1118 | 13.7 | 89.6 |
| | 3 | 674 | 8.3 | 97.8 |
| | 4 | 164 | 2.0 | 99.8 |
| | 5 | 14 | 0.2 | 100.0 |
| | all | 8161 | 100.0 | |
| PARENT1 | No | 7084 | 86.8 | 86.8 |
| | Yes | 1077 | 13.2 | 100.0 |
| | all | 8161 | 100.0 | |
| MSTATUS | No | 3267 | 40.0 | 40.0 |
| | Yes | 4894 | 60.0 | 100.0 |
| | all | 8161 | 100.0 | |
| SEX | F | 4375 | 53.6 | 53.6 |
| | M | 3786 | 46.4 | 100.0 |
| | all | 8161 | 100.0 | |
| EDUCATION | Less Than High School | 1203 | 14.7 | 14.7 |
| | High School | 2330 | 28.6 | 43.3 |
| | Bachelors | 2242 | 27.5 | 70.8 |
| | Masters | 1658 | 20.3 | 91.1 |
| | PhD | 728 | 8.9 | 100.0 |
| | all | 8161 | 100.0 | |
| JOB | | 526 | 6.4 | 6.4 |
| | Blue Collar | 1825 | 22.4 | 28.8 |
| | Clerical | 1271 | 15.6 | 44.4 |
| | Doctor | 246 | 3.0 | 47.4 |
| | Home Maker | 641 | 7.8 | 55.2 |
| | Lawyer | 835 | 10.2 | 65.5 |
| | Manager | 988 | 12.1 | 77.6 |
| | Professional | 1117 | 13.7 | 91.3 |
| | Student | 712 | 8.7 | 100.0 |
| | all | 8161 | 100.0 | |
| CAR_USE | Commercial | 3029 | 37.1 | 37.1 |
| | Private | 5132 | 62.9 | 100.0 |
| | all | 8161 | 100.0 | |
| CAR_TYPE | Minivan | 2145 | 26.3 | 26.3 |
| | Panel Truck | 676 | 8.3 | 34.6 |
| | Pickup | 1389 | 17.0 | 51.6 |
| | Sports Car | 907 | 11.1 | 62.7 |
| | SUV | 2294 | 28.1 | 90.8 |

Table 2: Table of nominal variables

| Variable | Levels | n | % | ∑ % |
|---|---|---|---|---|
| | Van | 750 | 9.2 | 100.0 |
| | all | 8161 | 100.0 | |
| RED_CAR | no | 5783 | 70.9 | 70.9 |
| | yes | 2378 | 29.1 | 100.0 |
| | all | 8161 | 100.0 | |
| CLM_FREQ | 0 | 5009 | 61.4 | 61.4 |
| | 1 | 997 | 12.2 | 73.6 |
| | 2 | 1171 | 14.3 | 88.0 |
| | 3 | 776 | 9.5 | 97.5 |
| | 4 | 190 | 2.3 | 99.8 |
| | 5 | 18 | 0.2 | 100.0 |
| | all | 8161 | 100.0 | |
| REVOKED | No | 7161 | 87.8 | 87.8 |
| | Yes | 1000 | 12.2 | 100.0 |
| | all | 8161 | 100.0 | |
| URBANICITY | Highly Rural/ Rural | 1669 | 20.4 | 20.4 |
| | Highly Urban/ Urban | 6492 | 79.5 | 100.0 |
| | all | 8161 | 100.0 | |

### 3.1.2 Continuous and Discrete Variables

We can see that in our continuous and discrete variables there is some additional variability. The median claim amount (`TARGET_AMT`) is 0 which would coincide with only a quarter for records indicating an accident. However, the spread is large since the average payout is only $1,504.30 but the maximum payout was $107,586.10. Surprisingly, the median `AGE` is 45 and the average `AGE` is 44.8 years, while we expected a lower average it could be due to simple selection bias in the data set source or the aging US population bringing this average higher [2]. We also noticed that an `INCOME` of $0.00 seems unwise because it is unclear how the individual would be able to cover their premium costs without parental support. Finally, we should note that the data set has as `CAR_AGE` of -3, which is impossible and will need to be removed.

There are many missing values for this portion of our data set, we have over 400 values missing for years on the job, income, home value, and car age. Due to these missing values we will need to impute to complete our statistical model.

| Variable | n | Min | $q_1$ | $\tilde{x}$ | $\bar{x}$ | $q_3$ | Max | s | IQR | #NA |
|---|---|---|---|---|---|---|---|---|---|---|
| TARGET_AMT | 8161 | 0 | 0 | 0 | 1504.3 | 1036 | 107586.1 | 4704.0 | 1036 | 0 |
| TIF | 8161 | 1 | 1 | 4 | 5.4 | 7 | 25.0 | 4.1 | 6 | 0 |
| AGE | 8155 | 16 | 39 | 45 | 44.8 | 51 | 81.0 | 8.6 | 12 | 6 |
| YOJ | 7707 | 0 | 9 | 11 | 10.5 | 13 | 23.0 | 4.1 | 4 | 454 |
| INCOME | 7716 | 0 | 28097 | 54028 | 61898.1 | 85986 | 367030.0 | 47572.7 | 57889 | 445 |
| HOME_VAL | 7697 | 0 | 0 | 161160 | 154867.3 | 238724 | 885282.0 | 129123.8 | 238724 | 464 |
| TRAVTIME | 8161 | 5 | 22 | 33 | 33.5 | 44 | 142.0 | 15.9 | 22 | 0 |
| BLUEBOOK | 8161 | 1500 | 9280 | 14440 | 15709.9 | 20850 | 69740.0 | 8419.7 | 11570 | 0 |
| OLDCLAIM | 8161 | 0 | 0 | 0 | 4037.1 | 4636 | 57037.0 | 8777.1 | 4636 | 0 |
| MVR_PTS | 8161 | 0 | 0 | 1 | 1.7 | 3 | 13.0 | 2.1 | 3 | 0 |
| CAR_AGE | 7651 | -3 | 1 | 8 | 8.3 | 12 | 28.0 | 5.7 | 11 | 510 |

Table 3:

[2] Ortman, Jennifer M., Victoria A. Velkoff, and Howard Hogan. "An aging nation: the older population in the United States." Washington, DC: US Census Bureau (2014): 25-1140.

## 3.2 Imputting Missing Values

In order to address the missing values in our variables we used a non-parametric imputation method (Random Forest) using the `missForest` package. The function is particularly useful in that it can handle any type of input data and it will make as few assumptions about the structure of the data as possible.[3]

**Table 2 : Imputed Descriptive Statistics**
**25 Variables 8161 Observations**

**TARGET_FLAG**

| n | missing | distinct | Info | Sum | Mean | Gmd |
|---|---|---|---|---|---|---|
| 8161 | 0 | 2 | 0.583 | 2153 | 0.3 | 0.4 |

**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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8161 | 0 | 66 | 0.999 | 45 | 10 | 30 | 33 | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8161 | 0 | 442 | 0.991 | 10 | 4 | 0 | 5 | 9 | 11 | 13 | 14 | 15 |

```
lowest :  0.00  0.11  0.14  0.24  0.30, highest: 16.00 17.00 18.00 19.00 23.00
```

**INCOME**

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8161 | 0 | 7057 | 1 | 61578 | 50857 | 0e+00 | 5e+03 | 3e+04 | 5e+04 | 9e+04 | 1e+05 | 2e+05 |

```
lowest : -9.255018e-11  0.000000e+00  5.000000e+00  7.000000e+00  1.545000e+01
highest:  3.062770e+05  3.096280e+05  3.201270e+05  3.323390e+05  3.670300e+05
```

**PARENT1**

| n | missing | distinct |
|---|---|---|
| 8161 | 0 | 2 |

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

**HOME_VAL**

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8161 | 0 | 5570 | 0.978 | 2e+05 | 1e+05 | 0e+00 | 0e+00 | 0e+00 | 2e+05 | 2e+05 | 3e+05 | 4e+05 |

```
lowest :      0.000   5427.070   7417.370   7553.793   8509.623
highest: 657804.000 682634.000 738153.000 750455.000 885282.000
```

[3] Stekhoven, Daniel J., and Peter Bühlmann. "MissForest-non-parametric missing value imputation for mixed-type data." Bioinformatics 28.1 (2012): 112-118.

### MSTATUS

| n | missing | distinct |
|---|---|---|
| 8161 | 0 | 2 |

No (3267, 0.4), Yes (4894, 0.6)

### SEX

| n | missing | distinct |
|---|---|---|
| 8161 | 0 | 2 |

F (4375, 0.536), M (3786, 0.464)

### EDUCATION

| n | missing | distinct |
|---|---|---|
| 8161 | 0 | 5 |

lowest : Bachelors High School Less Than High School Masters PhD
highest: Bachelors High School Less Than High School Masters PhD

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

### JOB

| n | missing | distinct |
|---|---|---|
| 8161 | 0 | 9 |

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

| Value | Blue Collar | Clerical | Doctor | Home Maker | Lawyer |
|---|---|---|---|---|---|
| Frequency | 526 | 1825 | 1271 | 246 | 641 | 835 |
| Proportion | 0.064 | 0.224 | 0.156 | 0.030 | 0.079 | 0.102 |

| Value | Manager | Professional | Student |
|---|---|---|---|
| Frequency | 988 | 1117 | 712 |
| Proportion | 0.121 | 0.137 | 0.087 |

### 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 | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8161 | 0 | 2789 | 1 | 15710 | 9354 | 4900 | 6000 | 9280 | 14440 | 20850 | 27460 | 31110 |

lowest : 1500 1520 1530 1540 1590, highest: 57970 61050 62240 65970 69740

### 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 SUV
highest: Panel Truck Pickup Sports Car SUV Van

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

### RED_CAR

| n | missing | distinct |
|---|---|---|
| 8161 | 0 | 2 |

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

---

### OLDCLAIM

```
     n    missing   distinct       Info       Mean        Gmd        .05        .10        .25        .50        .75        .90        .95
  8161          0       2857      0.769       4037       6563          0          0          0          0       4636       9583      27090

lowest :     0   502   506   518   519, highest: 52507 53477 53568 53986 57037
```

---

### 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
  8161          0        512      0.985          8          6          1          1          4          8         12         16         18

lowest : -3.000000  0.000000  1.000000  2.000000  2.382143
highest: 24.000000 25.000000 26.000000 27.000000 28.000000
```

---

### URBANICITY

```
     n    missing   distinct
  8161          0          2

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

---

## 3.3 Exploration of Variables

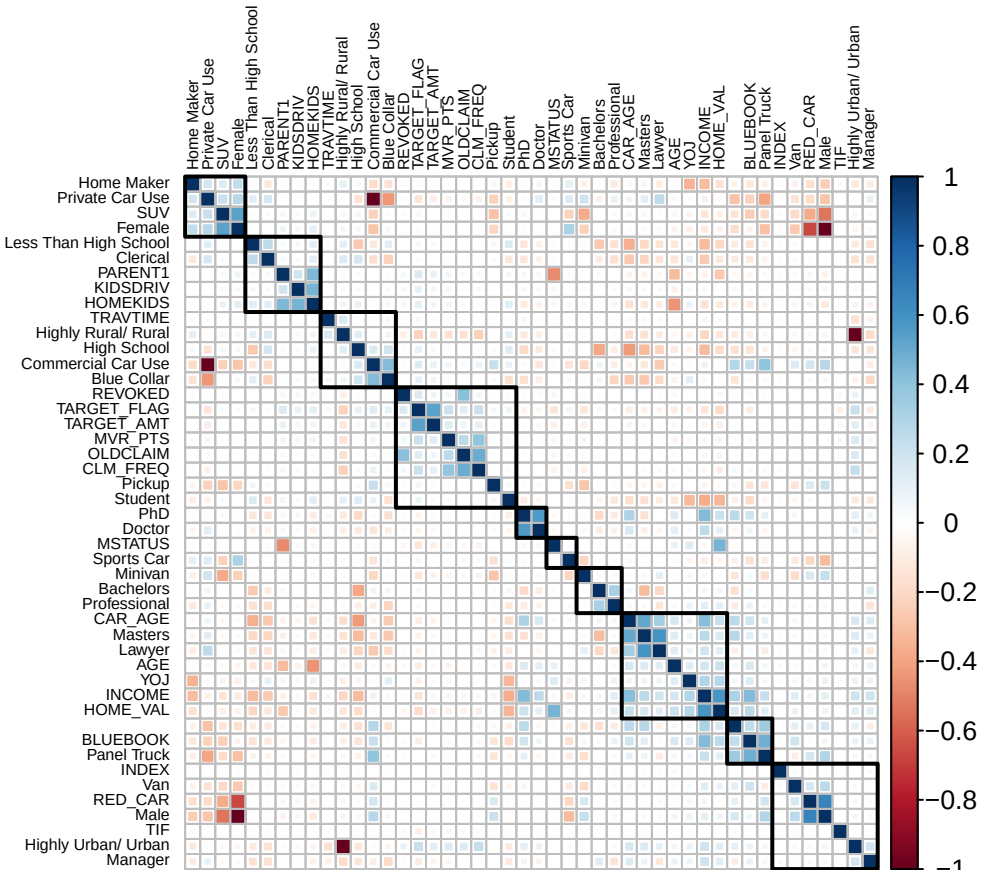


Figure 1: Correlation Plot of Training Data Set