A Capstone project report on

Risk Analytics model to understand the claim propensity for a car insurance company

Submitted by

Rithu A Mantagani

Submitted towards the partial fulfilment for the award of the degree in

Business Analytics

Submitted to:

Great Lakes Institute of Management

Post Graduate Program in Business Analytics and Business Intelligence

TABLE OF CONTENTS

List of Figures

List of tables

Introduction

1.1 Abstract

As of 2016, there were over 268 million registered vehicles on the roads in the United States. In 2015, there were 32,166 fatalities, 1,715,000 injuries and 4,548,000 car crashes which involved property damage. So, while many of us feel secure in our vehicles, the statistics indicate the importance of automobile insurance, and in most cases, auto insurance is required by law. Auto insurance is essential because it covers not only any physical damage that may occur in an accident but also any damage or injury that might be caused because of a vehicular accident or which may be done upon oneself or one's vehicle by another vehicle or accident – a falling tree for example.

The Insurance industry is immensely Data-intensive. Historically, there data has been mostly fragmented and underutilized. The insurance industry goes through much combining structured and unstructured data, which enables the sector to generate powerful insights. With an incredible amount of data flowing in from multiple new digital channels, the insurance industry is undergoing a paradigm shift in the way they function – right from product planning to pricing, introduction, marketing, customer self-service and claim processing.

1.2 Scope of the project:

The objective of this project is to build a "Risk Analytics model" to understand the renewal potential and claim propensity of Existing Customers under Personal Auto Insurance company.

1.3 Problem Statement:

Will the policyholder initiate a claim (Yes/No) for this policy in the next policy year?

1.4 Available Artefacts

Please find below-attached artefacts available concerning data and problem statement which were given to us.

11. Car Insurance\
Car insurance data s

PRE-PROCESSING

2.1 Data Description:

The dataset is a primary source of data collected during the year 2015-2016 for an auto insurance company. Cars manufactured through the year 1957-2000 are recorded in this data set.

It consists of 127 features and record for 14178 policies.

11. Car Insurance\
Car insurance data s

Table 1. Data description

Feature Name	Feature Type	Feature Description
Claim Status	Nominal	Indicates whether the policyholder has made a claim or not. 1 shows a claim and 0 indicates no claim
Claim Frequency	Ordinal	Gives the number of claims claimed
Premium	Numeric	Premium in $1000
Billing Term	Ordinal	How often the premium is paid, i.e., once a year = 1, Three times in a year = 3, or 6 times in a year = 6
Renewed	Nominal	Indicates whether the policy has been renewed or not. 1= Renewed, 0 = not renewed
DOB1	Date	Date of Birth of the central policyholder/ main driver
DOB2	Date	Date of Birth of the second driver
DOB3	Date	Date of Birth of the third driver
DOB4	Date	Date of Birth of the fourth driver
DOB5	Date	Date of Birth of the fifth driver
Number_of_Driver	Nominal	Count of the number of drivers in the policy
AgeUSdriving_1	Numeric	How long the driver has been driving

AgeUSdriving_2	Numeric	How long the driver has been driving
AgeUSdriving_3	Numeric	How long the driver has been driving
AgeUSdriving_4	Numeric	How long the driver has been driving
AgeUSdriving_5	Numeric	How long the driver has been driving
Amendment	Nominal	Number of changes made to the policy during the year
Coverage Liability	Nominal	The three numbers represent (in the $ thousands) the liability limits for per-person bodily injury, bodily injury for all persons injured in any one accident, and property damage liability. For example, say you live in Ohio and hold the minimum amount of coverage, which is 25/50/25. This means that the minimum liability limits in this state are $25,000 for injuries to one person, $50,000 for all injuries incurred and $25,000 for property damage for one vehicle in an accident.
Coverage MP	Nominal	Coverage M.P. stands for Medical payments coverage pays the reasonable expenses an insured person incurs for medical and funeral services within three years of an accident.
CoveragePD_1	Nominal	CoveragePD_1 stands for Property Damage. It is a type of liability coverage Property damage insurance covers any damages to someone's property. For example, the policy might show that you have Property Damage coverage of $25,000 per property, with a maximum of $50,000 per accident.
CoveragePIP_CDW	Nominal	Personal Injury Protection (PIP)-This a package of first-party medical benefits that provides broad protection for medical costs, lost wages, loss of essential services usually offered by the injured person (i.e., childcare, housekeeping), and funeral costs.
Coverage UMBI	Nominal	Uninsured/Underinsured motorist bodily injury
CoverageUMPD	Nominal	Uninsured/Underinsured motorist property damage coverage
DistanceToWork_1	Numeric	Distance to work for the first driver
DistanceToWork_2	Numeric	Distance to work for the second driver

DistanceToWork_3	Numeric	Distance to work for the third driver
DistanceToWork_4	Numeric	Distance to work for the fourth driver
DistanceToWork_5	Numeric	Distance to work for the fifth driver
DriverAssigned_1	Nominal	Count of drivers assigned to the first vehicle. 1 to 5
Engine_1	Nominal	Engine specification size in litres for the first vehicle
ExcludedDriverName_01	Nominal	First-person declared as an excluded driver
ExcludedDriverName_02	Nominal	The second person declared as an excluded driver
ExcludedDriverName_03	Nominal	Third-person declared as an excluded driver
ExcludedDriverName_04	Nominal	The fourth person declared as an excluded driver
ExcludedDriverName_05	Nominal	Fifth person declared as an excluded driver
ExcludedDriverName_06	Nominal	Sixth person declared as an excluded driver
ExcludedDriverName_07	Nominal	Seventh person declared as an excluded driver
ExcludedDriverName_08	Nominal	Eighth person declared as an excluded driver
ExcludedDriverName_09	Nominal	Ninth person declared as an excluded driver
ExcludedDriverName_10	Nominal	Tenth person declared as an excluded driver
ExcludedDriverName_11	Nominal	Eleventh person declared as an excluded driver
ExcludedDriverName_12	Nominal	Twelfth person declared as an excluded driver
ExcludedDriverName_13	Nominal	Thirteenth person declared as an excluded driver
ExcludedDriverName_14	Nominal	Fourteenth person declared as an excluded driver
ExcludedDriverName_15	Nominal	Fifteenth person declared as an excluded driver
ExcludedDriverName_16	Nominal	Sixteenth person declared as an excluded driver
ExcludedDriverName_17	Nominal	Seventeenth person declared as an excluded driver
ExcludedDriverName_18	Nominal	Eighteenth person declared as an excluded driver
ExcludedDriverName_19	Nominal	Nineteenth person declared as an excluded driver
ExcludedDriverName_20	Nominal	Twentieth person declared as an excluded driver
GaragedZIP_1	Nominal	Zip code of the place where the first vehicle is parked.
MaritalStatus_1	Nominal	Marital Status of the first driver. M - Married or S – Single

MaritalStatus_2	Nominal	Marital Status of the second driver. M - Married or S – Single
MaritalStatus_3	Nominal	Marital Status of the third driver. M - Married or S – Single
MaritalStatus_4	Nominal	Marital Status of the fourth driver. M - Married or S – Single
MaritalStatus_5	Nominal	Marital Status of the fifth driver. M - Married or S – Single
Occupation_1	Nominal	Occupation of the first driver
Occupation_2	Nominal	Occupation of the second driver
Occupation_3	Nominal	Occupation of the third driver
Occupation_4	Nominal	Occupation of the fourth driver
Occupation_5	Nominal	Occupation of the fifth driver
Relation_1	Nominal	Relationship of the first driver with the primary policyholder. Only Self
Relation_2	Nominal	Relationship of the second driver with the primary policyholder
Relation_3	Nominal	Relationship of the third driver with the primary policyholder
Relation_4	Nominal	Relationship of the fourth driver with the primary policyholder
Relation_5	Nominal	Relationship of the fifth driver with the primary policyholder
Rental_1	Nominal	first vehicle (If a rental is allowed)
Sex_1	Nominal	Gender of the first driver M - Male, F - Female
Sex_2	Nominal	Gender of the second driver M - Male, F - Female
Sex_3	Nominal	Gender of the third driver M - Male, F - Female
Sex_4	Nominal	Gender of the fourth driver M - Male, F - Female
Sex_5	Nominal	Gender of the fifth driver M - Male, F - Female
Surcharge1Unit_1	Nominal	First surcharge for the first vehicle. Y - Yes, N- No
Surcharge2Unit_1	Nominal	Second surcharge for the first vehicle Y - Yes, N- No
Surcharge3Unit_1	Nominal	Third surcharge for the first vehicle Y - Yes, N- No

Towing_1	Nominal	protects ones against some of the costs and hassles associated with frequent roadside breakdowns like dead batteries, flat tires or even an embarrassing lockout
Units	Nominal	Number of vehicles covered in the policy
VehicleInspected_1	Nominal	the first vehicle inspected. 1 - Vehicle was inspected, 0 - Vehicle was not inspected
ViolPoints1Driver_1	Nominal	First time the first driver is scoring a violation point.
ViolPoints1Driver_2	Nominal	First time the second driver is scoring a violation point.
ViolPoints1Driver_3	Nominal	First time the third driver is scoring a violation point.
ViolPoints1Driver_4	Nominal	First time the fourth driver is scoring a violation point.
ViolPoints1Driver_5	Nominal	First time the fifth driver is scoring a violation point.
ViolPoints2Driver_1	Nominal	Second time the first driver is scoring a violation point.
ViolPoints2Driver_2	Nominal	Second time the second driver is scoring a violation point.
ViolPoints2Driver_3	Nominal	Second time the third driver is scoring a violation point.
ViolPoints2Driver_4	Nominal	Second time the fourth driver is scoring a violation point.
ViolPoints2Driver_5	Nominal	Second time the fifth driver is scoring a violation point.
ViolPoints3Driver_1	Nominal	Third time the first driver is scoring a violation point.
ViolPoints3Driver_2	Nominal	Third time the second driver is scoring a violation point.
ViolPoints3Driver_3	Nominal	Third time the third driver is scoring a violation point.
ViolPoints3Driver_4	Nominal	Third time the fourth driver is scoring a violation point.
ViolPoints3Driver_5	Nominal	Third time the fifth driver is scoring a violation point.
ViolPoints4Driver_1	Nominal	Fourth time the first driver is scoring a violation point.
ViolPoints4Driver_2	Nominal	Fourth time the second driver is scoring a violation point.
ViolPoints4Driver_3	Nominal	Fourth time the third driver is scoring a violation point.
ViolPoints4Driver_4	Nominal	Fourth time the fourth driver is scoring a violation point.

ViolPoints4Driver_5	Nominal	Fourth time the fifth driver is scoring a violation point.
ViolPoints5Driver_1	Nominal	Fifth time the first driver is scoring a violation point.
ViolPoints5Driver_2	Nominal	Fifth time the second driver is scoring a violation point.
ViolPoints5Driver_3	Nominal	Fifth time the third driver is scoring a violation point.
ViolPoints5Driver_4	Nominal	Fifth time the fourth driver is scoring a violation point.
ViolPoints5Driver_5	Nominal	Fifth time the fifth driver is scoring a violation point.
ViolPoints6Driver_1	Nominal	Sixth time the first driver is scoring a violation point.
ViolPoints6Driver_2	Nominal	Sixth time the second driver is scoring a violation point.
ViolPoints6Driver_3	Nominal	Sixth time the third driver is scoring a violation point.
ViolPoints6Driver_4	Nominal	Sixth time the fourth driver is scoring a violation point.
ViolPoints6Driver_5	Nominal	Sixth time the fifth driver is scoring a violation point.
ViolPoints7Driver_1	Nominal	Seventh time the first driver is scoring a violation point.
ViolPoints7Driver_2	Nominal	Seventh time the second driver is scoring a violation point.
ViolPoints7Driver_3	Nominal	Seventh time the third driver is scoring a violation point.
ViolPoints7Driver_4	Nominal	Seventh time the fourth driver is scoring a violation point.
ViolPoints7Driver_5	Nominal	Seventh time the fifth driver is scoring a violation point.
ViolPoints8Driver_1	Nominal	Eighth time the first driver is scoring a violation point.
ViolPoints8Driver_2	Nominal	Eighth time the second driver is scoring a violation point.
ViolPoints8Driver_3	Nominal	Eighth time the third driver is scoring a violation point.
ViolPoints8Driver_4	Nominal	Eighth time the fourth driver is scoring a violation point.
ViolPoints8Driver_5	Nominal	Eighth time the fifth driver is scoring a violation point.
Year_1	Nominal	Year of manufacture of the first vehicle
Make_1	Nominal	Make of the first vehicle
Model_1	Nominal	Model of the first vehicle

Zip	Nominal	Zip code
Total_Distance_To_Work	Numeric	Total Distance to work of all the drivers combined
No Loss Signed	Nominal	Whether a statement of No loss has been signed or not. 1 - yes and 0 – No
Type	Nominal	Different types of auto insurance viz, A, AA.P. DD.P. FF.C. P, REN, RET, VV.D. XFR
Cancellation Type	Nominal	Type of cancellation viz, NN.P. INS

2.2 Data Cleaning:

Please find below the detailed account of the data processing done on all of the data.

Functional_mapping.xlsx

2.2.1 Removal of unwanted variables:

Some variables had to be removed at the beginning of analysis because:

- The variable like violation Point 1 till violation point 8 for driver 2 to driver five were having more than 70% missing values.
- Model_1 had around 354 levels.
- Occupation1 to Occupation 5 were having more than 400 to 1200 levels.
- Engine_1 contains incorrect/unknown format which cannot be transformed into one single type/format.
- When talking about dataset for no of drivers=1 then we had to drop the driver information corresponding to rest of the four drivers. This approach is taken for the rest of the datasets.
- Relation_1 since we know Relation_1 has only 'self' as a level in the data frame. Such information does not help us classify/differentiate records.
- Coverage UMPD behaviour is same as Coverage UMBI column. Since this column is not adding any more information, we are dropping this column.
- Distance_to_work_1-5 is dropped because total Distance to work attribute is helping us segregate claims better.
- Claim Frequency has a perfect co-relation with claim status. If a claim has been made once or more times, obviously claim status becomes one moreover, this information cannot be known before policy tenure ends; hence this is moved out of scope.
- We had discovered that all the zip codes belonged to the state of Texas. Since there are too many levels in the variable, it was dropped. However, zip

code/location-based analysis requires more domain-based insightful transformation for usage.

- MaritalStatus_3,4,5: These variables were dropped since the enrichment in them was less than 95%.
- Occupation_(x): Occupation of the five declared drivers and too many levels and data enrichment were low. Hence these were dropped.

2.2.2 Missing Value Treatment:

- Most of the variables have had missing values.
- If a column has missing/null values more than 10%, then we have removed the column from our analysis. This was done to avoid the generation of false trends.
- Variables with missing values were handled by using MICE package / Median / Mode techniques in R.

2.2.3 Outlier treatment:

- Few glitches found in the data, namely in Towing_1, Year_1, Rental_1 column which was treated with Median/Mode techniques.
- Since premium variable was found with many values on the right to 3rd quartile, it is assumed to be the trend of data in that variable and that there are no data glitches.
- Make_1 variable was text mined for accuracy in the brand information of the first insured vehicle in the policy.

2.2.4 Text Mining:

- Make_1 variable had around 90 levels which got transformed into 40 levels.
- Further, Make_1 was binned as either popular brand or "others."

2.2.5 Clubbing of levels

- A variable has more than two levels with an imbalanced number of records and claim rates.
- Binning makes logical sense, and if binning is giving us better distinctive claim rates among classes.

Some of the examples include:

1. Billing term

```
> table(df_data_4$Billing_Term,df_data_4$ClaimStatus)

    0   1
1  38   6
3   6   0
6 107  27
> df_data_4$Billing_Term=ifelse(df_data_4$Billing_Term==3 | df_data_4$Billing_Term==6 , 2, 1)
> table(df_data_4$Billing_Term,df_data_4$ClaimStatus)

    0   1
1  38   6
2 113  27
```

Fig1. Bins of "Billing term."

2. Amendments

```
> table(df_data_4$Amendment,df_data_4$ClaimStatus)

      0   1
0   130  30
1     8   0
2    10   0
3     2   0
4     1   1
7     0   1
10    0   1
> df_data_4$Amendment=ifelse(df_data_4$Amendment==0 , 0, 1)
> table(df_data_4$Amendment,df_data_4$ClaimStatus)

      0   1
0   130  30
1    21   3
```

Fig2. Bins of "Amendments"

Similarly, "Coverage Liability", "DriverAssigned_1", "Year_1", "total_voilation_points" and "MaritalStatus_2" there are many other instances across the four models where binning was done based on similar logic.

2.2.6 Addition of New variables:

- Excluded drivers in the policies could be a maximum of 20 in number. This information was given in 20 columns as the name of such drivers. We had then taken a total count of such instances in once policy to say what the total number of excluded drivers were.

2.3 Exploratory Data Analytics:

- The ratio of overall claims in the dataset is 5.48%
- We have split the dataset into four subsets based on the number of drivers registered for policy.
- To avoid overfitting, we have decided to model the problem statement (identifying the customer that will claim) by splitting/sub setting data by the number of drivers declared in the policy. The related irrelevant variables for each scenario were dropped and modified accordingly.

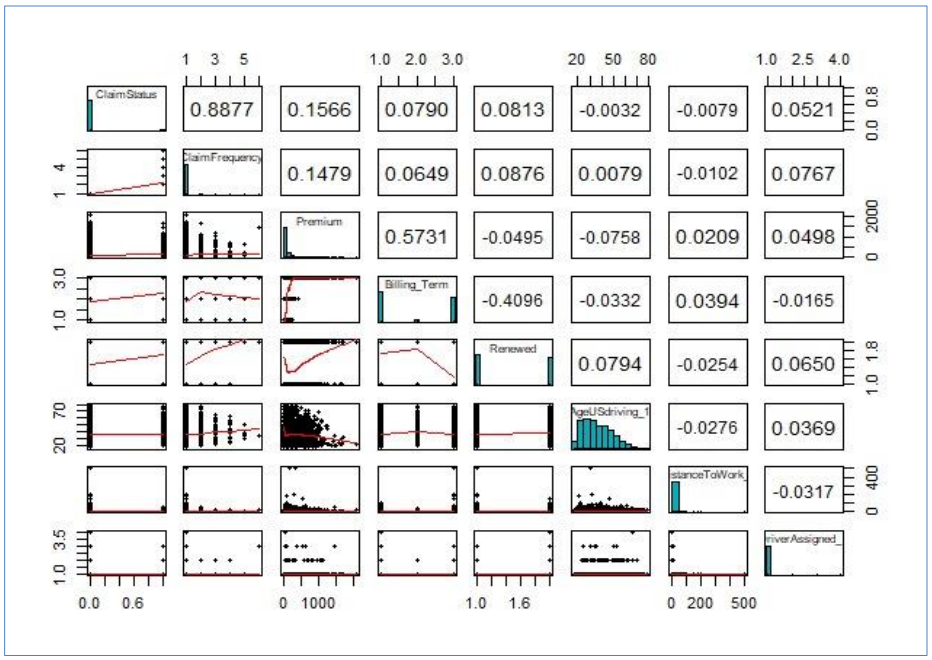
  -Number of Drivers equal to 1

  -Number of Drivers equal to 2

  -Number of Drivers equal to 3

-Number of Drivers equal to 4 or 5

- Correlation plot for variables vs claim status



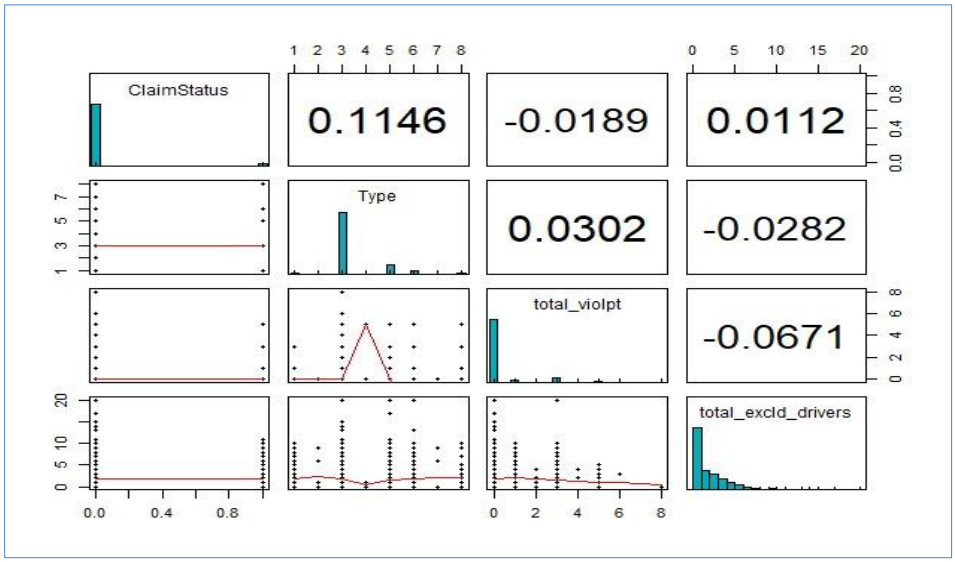


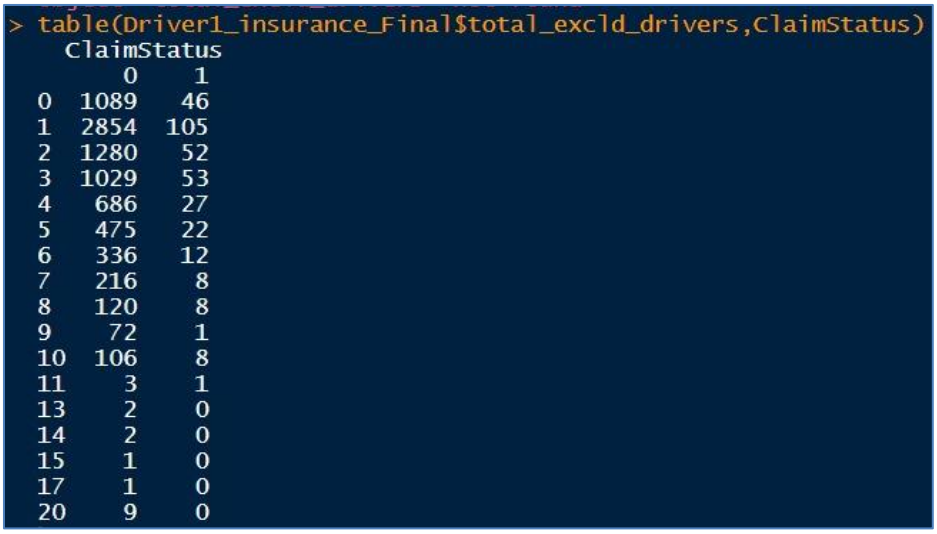
Fig3. Correlation plot with Predictor variables

- Policies which have billing terms less than or equal to 3 tend to be claimed more.

```
> table(ClaimStatus, Billing_Term)
           Billing_Term
ClaimStatus    1    3    6
          0 4487  300 3494
          1  109   31  203
```

Fig4. Distribution of Claim Status against Billing term

- Looking at the total excluded drivers, we see that the claims tend to increase and peak at one excluded driver and then decrease over excluded driver counts in the policy.



```
> table(Driver1_insurance_Final$total_excld_drivers,ClaimStatus)
   ClaimStatus
       0    1
  0 1089   46
  1 2854  105
  2 1280   52
  3 1029   53
  4  686   27
  5  475   22
  6  336   12
  7  216    8
  8  120    8
  9   72    1
  10 106    8
  11   3    1
  13   2    0
  14   2    0
  15   1    0
  17   1    0
  20   9    0
```

Fig5. Claim Status against Excluded Drivers

2.4 Feature engineering, if necessary:

The data (spread across 127 columns) had multiple features for a policy relating to multiple drivers declared in the insurance plans, enabling much scope for feature engineering and selection for model building. We have tried the following rules for feature selection and feature engineering for all models:

- Date transformation was also done to impute missing values in Age_US_Driving variables for the five drivers. Taking an example in model 4,
    a) The five dob values and age us driving costs are assumed to give out the same information.
    b) The age_us_driving_1 and dob1 variable give out the age of the first driver declared the policy. Using this, we have computed a reference year by adding the year of dob to age_us driving one. This gave us a reference year.
    c) Using this reference year, we have subtracted the dob2, dob3, dob4, dob5 values to extract the age of drivers where the data was missing.

- **Total_Distance_To_Work**: Total Distance to work variable had a lot of zero vales and was heavily skewed to the right. In these conditions, we applied the log10(x+1) transformation for better distribution. PDF distribution before and after transformation:

Before **After**

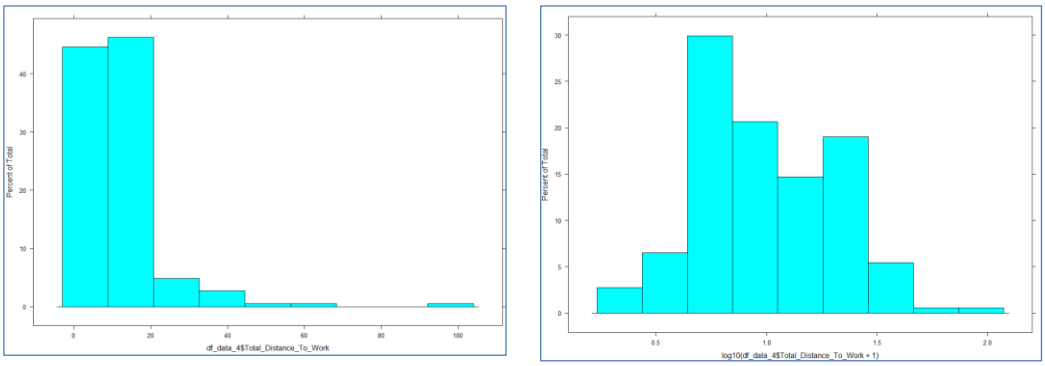


Fig6. Data skewness for "Total Distance to work."

- **Premium:** Total Distance to work variable had a lot of zero vales and was heavily skewed to the right. In these conditions, we applied the log10(x+1) transformation for better distribution. The before and after distribution looks more or less the same as the total Distance to work transformation above.

Model Building

3.1 Data split: Training & Test

The data was split 75-25 or 70-30 as train and Test SMOTE was applied to balance out Claim Status classes. Please find below before and after claim ratios of train data and test data. We used SMOTE on train data.

Data Split	Train		Test		Train-after SMOTE	
	Y	N	Y	N	Y	N
Model 1	3.75%	96.25%	4.2%	92.72%	43.75%	56.25%
Model 2	6.96%	93.04%	6.91%	93.09%	60%	40%
Model 3	10.9%	89.1%	10.9%	89.1%	50.1%	49.9%
Model 4	17.99%	82.01%	17.77%	82.23%	50%	50%

Table2. Train and test for four split data set

3.2 Modelling Process

This section describes the modelling process undertaken to make a predictive model for Claim of car insurance.

3.2.1 Modelling approach and Validation

- We commenced the modelling by considering positive class as '1' and built a CART decision tree.
- The high accuracy and low sensitivity in the initial CART model were interpreted to be a model with a bias towards the majority class of claim status variable
- To remediate this problem, we applied one of the most common ways of handling imbalanced data: SMOTE (under-sampling and over-sampling technique).
- Below are the results of the CART model without SMOTE

TRAIN data **TEST data**

```
> confusionMatrix(predicted,
+                 train_balanced$ClaimStatus,
+                 mode="everything",positive='
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 258  48
         1 131 344

               Accuracy : 0.7708
                 95% CI : (0.7397, 0.7999)
    No Information Rate : 0.5019
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5412

 Mcnemar's Test P-Value : 8.845e-10

            Sensitivity : 0.8776
            Specificity : 0.6632
         Pos Pred Value : 0.7242
         Neg Pred Value : 0.8431
              Precision : 0.7242
                 Recall : 0.8776
                     F1 : 0.7935
             Prevalence : 0.5019
         Detection Rate : 0.4405
   Detection Prevalence : 0.6082
      Balanced Accuracy : 0.7704

       'Positive' Class : 1
```

```
> confusionMatrix(predicted,
+                 testTransformed$ClaimStatus,
+                 mode="everything",positive='1')
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 125   9
         1  70  15

               Accuracy : 0.6393
                 95% CI : (0.5718, 0.7029)
    No Information Rate : 0.8904
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1258

 Mcnemar's Test P-Value : 1.473e-11

            Sensitivity : 0.62500
            Specificity : 0.64103
         Pos Pred Value : 0.17647
         Neg Pred Value : 0.93284
              Precision : 0.17647
                 Recall : 0.62500
                     F1 : 0.27523
             Prevalence : 0.10959
         Detection Rate : 0.06849
   Detection Prevalence : 0.38813
      Balanced Accuracy : 0.63301

       'Positive' Class : 1
```

Fig7. Confusion Matrix for CART model without SMOTE for Model 3

- After balancing the train data, we have applied a CART decision tree, Random Forest and multinomial logistic regression.
- Below are the results for Models using SMOTE Data.

	Train Data	Test data
CART	Confusion Matrix and Statistics Reference Prediction 0 1 0 305 100 1 84 292 Accuracy : 0.7644 95% CI : (0.733, 0.7938) No Information Rate : 0.5019 P-Value [Acc > NIR] : <2e-16 Kappa : 0.5289 Mcnemar's Test P-Value : 0.2688 Sensitivity : 0.7449 Specificity : 0.7841 Pos Pred Value : 0.7766 Neg Pred Value : 0.7531 Precision : 0.7766 Recall : 0.7449 F1 : 0.7604 Prevalence : 0.5019 Detection Rate : 0.3739 Detection Prevalence : 0.4814 Balanced Accuracy : 0.7645 'Positive' Class : 1	Confusion Matrix and Statistics Reference Prediction 0 1 0 133 9 1 62 15 Accuracy : 0.6758 95% CI : (0.6095, 0.7373) No Information Rate : 0.8904 P-Value [Acc > NIR] : 1 Kappa : 0.156 Mcnemar's Test P-Value : 6.775e-10 Sensitivity : 0.62500 Specificity : 0.68205 Pos Pred Value : 0.19481 Neg Pred Value : 0.93662 Precision : 0.19481 Recall : 0.62500 F1 : 0.29703 Prevalence : 0.10959 Detection Rate : 0.06849 Detection Prevalence : 0.35160 Balanced Accuracy : 0.65353 'Positive' Class : 1

Fig8. Confusion Matrix for CART model with SMOTE for Model 3

	Train Data	Test data
Random Forest	Confusion Matrix and Statistics Reference Prediction 0 1 0 361 14 1 28 378 Accuracy : 0.9462 95% CI : (0.928, 0.961) No Information Rate : 0.5019 P-Value [Acc > NIR] : < 2e-16 Kappa : 0.8924 Mcnemar's Test P-Value : 0.04486 Sensitivity : 0.9643 Specificity : 0.9280 Pos Pred Value : 0.9310 Neg Pred Value : 0.9627 Precision : 0.9310 Recall : 0.9643 F1 : 0.9474 Prevalence : 0.5019 Detection Rate : 0.4840 Detection Prevalence : 0.5198 Balanced Accuracy : 0.9462 'Positive' Class : 1	Confusion Matrix and Statistics Reference Prediction no yes no 145 13 yes 50 11 Accuracy : 0.7123 95% CI : (0.6475, 0.7713) No Information Rate : 0.8904 P-Value [Acc > NIR] : 1 Kappa : 0.1205 Mcnemar's Test P-Value : 5.745e-06 Sensitivity : 0.45833 Specificity : 0.74359 Pos Pred Value : 0.18033 Neg Pred Value : 0.91772 Precision : 0.18033 Recall : 0.45833 F1 : 0.25882 Prevalence : 0.10959 Detection Rate : 0.05023 Detection Prevalence : 0.27854 Balanced Accuracy : 0.60096 'Positive' Class : 1

Fig9. Confusion Matrix for RR.F.model with SMOTE for Model 3

	Train Data	Test data
Logistic Regression	Confusion Matrix and Statistics Reference Prediction 0 1 0 293 100 1 96 292 Accuracy : 0.749 95% CI : (0.7171, 0.7791) No Information Rate : 0.5019 P-Value [Acc > NIR] : <2e-16 Kappa : 0.4981 Mcnemar's Test P-Value : 0.8303 Sensitivity : 0.7449 Specificity : 0.7532 Pos Pred Value : 0.7526 Neg Pred Value : 0.7455 Precision : 0.7526 Recall : 0.7449 F1 : 0.7487 Prevalence : 0.5019 Detection Rate : 0.3739 Detection Prevalence : 0.4968 Balanced Accuracy : 0.7491 'Positive' Class : 1	Confusion Matrix and Statistics Reference Prediction 0 1 0 138 6 1 57 18 Accuracy : 0.7123 95% CI : (0.6475, 0.7713) No Information Rate : 0.8904 P-Value [Acc > NIR] : 1 Kappa : 0.2369 Mcnemar's Test P-Value : 2.988e-10 Sensitivity : 0.75000 Specificity : 0.70769 Pos Pred Value : 0.24000 Neg Pred Value : 0.95833 Precision : 0.24000 Recall : 0.75000 F1 : 0.36364 Prevalence : 0.10959 Detection Rate : 0.08219 Detection Prevalence : 0.34247 Balanced Accuracy : 0.72885 'Positive' Class : 1

Fig10. Confusion Matrix for Logistic Regression model with SMOTE for Model 3

- After Regularizing the models, it was observed that the logistic regression model was running most stable concerning differences between test and train parameters for all the four models.

- Please find Attached below recommended models for the four subsets of data named from model one till four.


final_models.xlsx

NOTE: Discussion on performance and interpretations of the best model are in the sections that follow.

3.2.2 Model Performance Metrics and Assumptions verification

3.2.2.1 Model significance test:

Model significance is checked using the log-likelihood test
Log-Likelihood Test: In statistics, a likelihood ratio test is a statistical test used for comparing the goodness of fit of two models, one of which (the null model) is a particular case of the other (the alternative model). The tests based on the likelihood ratio, which expresses how many times more likely, the data are under one model than the other. This likelihood ratio, or equivalently its logarithm, can then be used to compute a p-value, or compared to a critical value to decide whether to reject the null model and hence accept the alternative model.

```
> lrtest(logit1)
Likelihood ratio test

Model 1: ClaimStatus ~ Premium + Renewed + CoverageLiability + Amendment +
    Rental_1 + Units + Billing_Term + Towing_1 + MaritalStatus_1
Model 2: ClaimStatus ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  10 -418.72
2   1 -541.34 -9 245.25  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig11.Model Significant Check for Model 3

H0: All betas are zero
H1: At least one beta is nonzero

From the log-likelihood, we can see that, intercept only model -541.34 variance was unknown to us. When we take the full model, -418.72 variation was unknown to us.
So, we can say that 1-(-418.72 /-541.34) = 22.66% of the uncertainty inherent in the intercept-only model is calibrated by the full model.
Chisq likelihood ratio is significant. Also, the p-value suggests that we can accept the Alternate Hypothesis that at least one of the beta is not zero.
So, the model is significant.

3.2.2.2 Model robustness check:

Now since we concluded that the model built is significant, let's find out how robust it is with the help of McFadden pseudo-R Squared Test.

McFadden pseudo-R-square: McFadden pseudo-R Squared: Logistic regression models are fitted using the method of maximum likelihood. I.e. the parameter estimates are those values which maximize the likelihood of the Data which have been observed.

```
> pR2(logit1)
fitting null model for pseudo-r2
         llh      llhNull           G2     McFadden         r2ML         r2CU
-418.7155973 -541.3421862  245.2531777    0.2265232    0.2694989    0.3593336
```

Fig12.Model Robustness Check for Model 3

The McFadden's pseudo-R Squared TeTestuggests that at least 22.65% variance of the data is captured by our model, which suggests it's a robust model.

3.2.2.3 Heteroscedasticity check:

The solution to the regression problem becomes unstable in the presence of 2 or more correlated predictors. Multicollinearity can be measured by computing variance inflation factor (VIF) which gauges – how much the variance of the regression coefficient is inflated due to multicollinearity.

```
vif(logit1)
       Premium        Renewed CoverageLiability      Amendment       Rental_1
      1.338050       1.289448       1.132201       1.083984       1.305491
         Units   Billing_Term       Towing_1 MaritalStatus_1
      1.111105       1.497977       1.263958       1.058562
```

Fig13.Multicollinearity Check for Model 3

As a thumb rule, VIF of more than 5 or 10 is considered to be significant. And such variables can be removed to improve the stability of the regression model.

3.2.2.4 Receiver Operating Characteristic (ROC):

It is a plot of the True Positive Rate against the False Positive Rate for the different possible cut-points of a diagnostic test.
A ROC curve demonstrates several things:
• It shows the trade-off between sensitivity and specificity (a decrease in specificity will accompany any increase in sensitivity).

• The closer the curve follows the left-hand border and then the top edge of the ROC space, the more accurate the test

• The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test

• The slope of the tangent line at a cut-point gives the likelihood ratio (LL.R. for that value of the TeTestThe area under the curve (AUC) is a measure of text accuracy.
Accuracy is measured by the area under the ROC curve. An area of 1 represents a perfect test; an area of 0.5 represents a worthless test. A rough guide for classifying the accuracy of a model verification test is the traditional academic point system, as follows:
0.90-1 = excellent (A)
0.80-0.90 = good (B)
0.70-0.80 = fair (C)
0.60-0.70 = poor (D)
0.50-0.60 = fail (F)

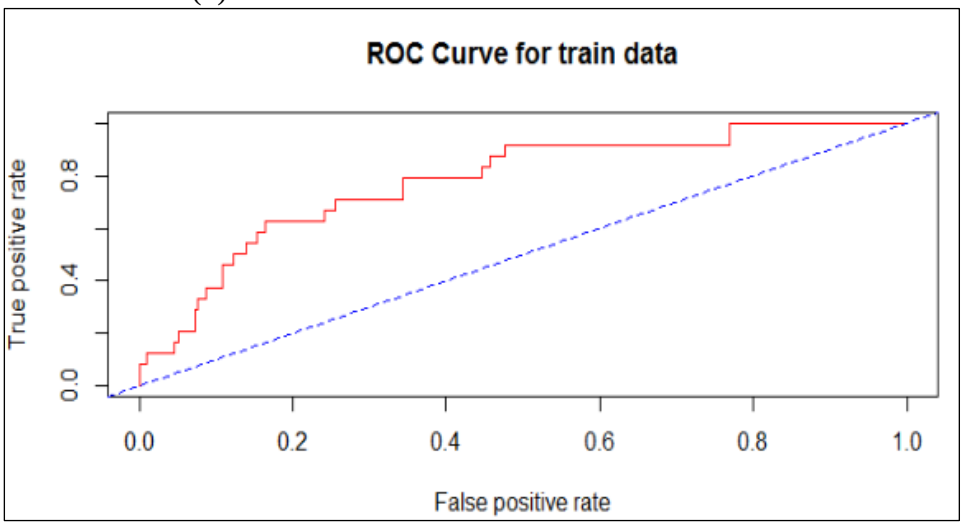


Fig14.ROC curve for Model 3

AUC At 0.7786, the ROC Curve of our model demonstrates reasonably good results.

3.2.2.5 The area under the curve (AUC):

AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive data point higher than a randomly chosen negative data point. Higher the chance better is the classifier.

```
> train.area
[1] 0.7910262
```

```
> test.area
[1] 0.7786325
```

Fig15.AUC curve for Model 3

3.2.2.6 Gini Coefficient:

Gini coefficient is a ratio of two areas:

- The area between the ROC curve and the random model line
- Top left triangle above the arbitrary model line – which is just 0.5

It can also be simplified as (2 * AUC – 1)

```
> train.gini
[1] 0.5820524
```

```
> test.gini
[1] 0.557265
```

Fig16.Gini Coefficient for Model 3

The loss for the model through GINI is 0.55

3.2.2.7 Kolmogorov–Smirnov test:

This performance measure is defined as the maximum difference between TPR and FPR. Higher KS stat value indicates a better model.

```
> ks.train
[1] 0.4934815
```

```
> test.ks
[1] 0.4608974
```

Fig16.K K.S.for Model 3

KS=46.08 indicates a good model.

Please find below model performance metrics for all four logistic regression models:

Models	McFadden R square value	AUC		GINI		KS	
		Train	Test	Train	Test	Train	Test
Model 1	36.01	82.28	75.86	64.56	51.7	64.56	51.7
Model 2	16.01	76.505	80.41	53.01	60.82	41.52	45.93
Model 4	22.05	73.83	71.54	51.04	49.03	51.04	49.03

Interpretation and Conclusion

4.1 Model Interpretation

Observing the recommended model for 3rd split of data, we observe:

LOGISTIC REGRESSION

Model 3

Estimate Std. Error z value Pr(>|z|)						Slope Values	Odds	Probability
(Intercept)	-5.342e+00	9.439e-01	-5.660	1.51e-08	***	-5.342		
Premium	1.373e-03	3.268e-04	4.201	2.66e-05	***	0.001373	1.001374	0.5003432
Billing_Term2	6.074e-01	2.560e-01	2.372	0.01767	*	0.6074	1.835652	0.6473475
Renewed1	1.333e+00	2.089e-01	6.380	1.77e-10	***	1.333	3.792404	0.7913364
Amendment1	5.226e-01	2.526e-01	2.069	0.03857	*	0.5226	1.686407	0.6277555
CoverageLiability1	1.155e+00	2.151e-01	5.370	7.86e-08	***	1.155	3.174023	0.760423
Rental_11	1.135e+00	4.814e-01	2.358	0.01835	*	1.135	3.111174	0.7567605
Towing_11	1.292e+00	5.143e-01	2.512	0.01200	*	1.292	3.640059	0.7844855
Units2	3.922e-01	1.919e-01	2.043	0.04102	*	0.3922	1.480234	0.5968122
Year_11	4.205e-01	2.001e-01	2.101	0.03561	*	0.4205	1.522723	0.6036029

- For one unit increase in log10(Premium) {since a log10 transformation was applied on Premium variable}, log odds of Claim Status being Y increases by 0.013
- If the policy is renewed, we observe that log odds of Claim Status being Y increases by 1.33
- A policy which has a billing term of frequency as three then log odds of Claim Status being Y increases by 0.6074
- If the policy has units more than one insured, log odds of Claim Status being Y increases by 0.3922
- Policies where amendments have been made, the log odds of Claim Status being Y increases by 0.5226
- Policies with coverage liability as '20/40/15' have the odds of Claim Status being Y increased by 1.155
- If the first cat insured under the policy is manufactured after the year 2000, the log odds of claim status being 'Y' increases by 0.4205
- If the first car insured under the policy is a rental allowed type vehicle, then the log odds of claims being made increase by 1.135
- If the first vehicle insured in the policy is covered for towing charges, then the log odds of claims being made increase by 1.292

4.2 Business insights

1. **Reducing Fraudulent Claims**: We can identify potentially fraudulent claims, allows for careful, extensive background checks.

2. **Improving Marketing by Forecasting Buyer Behavior**: Campaigns or promos can be arranged for the customer based on buyer behavior and increase the revenue for the insurance company.

3. **Triaging Claims**: Customers are always looking for fast, personalized service. By using this model, the insurance company will be able to prioritize specific claims to save time, money, and resources – needless to say, retain business and increase customer satisfaction.

4. **Identifying Outlier Claims**: Predictive analytics in insurance can help identify claims that unexpectedly become high-cost losses — often referred to as outlier claims. With proper analytics tools, insurers can review previous applications for similarities – and send alerts to claims specialists automatically. Advanced notice of potential losses or related complications can help insurers cut down on these outlier claims. A Machine learning model, especially multinomial logistic regression can help us identify when huge claims are made, and companies can access risk accordingly and restructure their offerings.

4.3 Final recommendation and thoughts on implementation

Since there were four models being recommended based on the number of drivers as 1,2,3 and 4-5, hence we have below recommendation for each policy segment:

1. Policies where the number of drivers is one:
    a. As there is a unit increase in the units insured, it is observed that the odds of a claim being made have increased significantly; hence the premium of policies where the number of drivers is one and units are more than one needs to be risk-adjusted.
    b. If the drivers assigned to the first vehicle are more than one, then the ABC firm should run the background check extensively for all the assigned drivers. More than one driver assigned to the first unit is a good indicator since the log odds of a claim being made is 1.62
    c. Discounts attributed to the renewal of a policy when the number of drivers is one should be minimized if any. The ABC firm should also consider providing them with good customer service. Even though they are more likely to make a claim, they will recommend the ABC firm's service to others being an old customer; it will have a bigger impact.

2. Policies where the number of drivers is two:
    a. If the age of the first insured driver is increased by ten years, then the odds of claiming policy increases by 3.3, hence tele markets should target younger customers in such a policy segment.
    b. The market segment where the marital status of the second driver is single, and the number of drivers is two, in that case, the odds of a claim being made decreases by 0.539, this can be used to run campaigns to target such customers, a study into the relationship between such drivers can be done to target such customers better.
3. Policies where the number of drivers is three:
    a. Policies where the first insured car is manufactured later than the year 2000, they tend to be claimed more. The market where the number of drivers is three, rental allowed and car manufactured after 2000, then these policies should be vetted carefully before confirming.
    b. Discounts attributed to the renewal of a policy when the number of drivers is three should be minimized if any. Telemarketers should not be keen on taking on this customer segment who are looking for a renewal of the policy with the company since odds of them making a claim is greater in this condition, and the liability on the ABC firm would be high in this segment.
4. Policies where the number of drivers is four:
    a. If Total excluded drivers in this segment is more than three, then the log odds of the policy being claimed is increased by 3.4. This customer segment should be appropriately vetted, and the premium of such policies should be risk-adjusted.
    b. When the number of drivers is five and the first driver declared is single, in such cases the log odds of a claim increase by 4.31 and hence such policies should be carefully risk-adjusted and vetted during onboarding.

4.4Conclusion

- The objective of identifying the policies which are likely to be claimed in this year has been achieved by using tools like R to recognize patterns and mitigate the financial risk of the insurance firm ABC accordingly.
- The study also helped identify the customer segments to focus on.
- The overall claim ratio from the live data would also give a ball park liability of the firm in the coming financial plan; hence budget planning can be done accordingly.
- The project work stands complete as far as identifying opportunities/risks and proposing solutions to the company's bottom line.

REFERENCES

1. Boodhun, N. (2017). A Review of Data Analytical Approaches in the Insurance Industry. *Journal of Applied Technology and Innovation, 1*(1), 58-73.
2. Study materials and videos from Great Lakes institute of management were referred in the preparation of this project.

DECLARATION

I hereby declare that the Project Report entitled "Risk Analytics model to understand the claim propensity for a car insurance company" has been submitted to Great Lakes Institute of Management for the PGP-BABI Certification. The work is original in its contents to the best of my knowledge and belief.

Date:3rd Jan 2021

Rithu A Mantagani