FINAL PROJECT

SUBMITTED BY

- ANOOP MADAMSETTY
- BATCH 15

PURPOSE

- What exactly is the objective?

  - The objective is to implement appropriate techniques which is taught in INSOFE and to get good insight of how to deal with real life projects.

  - To learn how data analytics enables quick decisions or help change policies due to trends observed.

  - To derive meaningful trends or intriguing findings that were not previously seen or empirically validated.

  - One of the reasons of choosing this project is to challenge myself to work on unknown domains and adapting of how to think and derive at conclusions of how to attack the problem to be solved.

  - In this project we have to find out the attributes such as growth rate and loss ratio which is used for improving the existing knowledge used for agent segmentation in a supervised predictive framework.

STATE-OF-THE-ART

- METRICS

  The differences between high growth agencies in their territory and those that were not growing or growing only marginally –

  a key differentiator for the high growth agencies is that they tracked and acted upon key metrics relating to the origins of their business, close ratios, revenue and policies per client, retention and average client tenure

  Agencies should track to maximize their marketing efforts and grow the business.

- GROWTH KEY METRICS :

  Let's take a look at what the high growth agencies specifically track to help achieve their high growth numbers. These tracking methods can help you grow too. The high growth agencies tracked 10 common items. They fall into four categories:

  - New business
  - Revenue
  - Retention/Renewal
  - Customer Service
  -

Revenue-It's Sub Metrics:

1. **Revenue per policyholder –**
   Revenue per Policyholder is a simple key performance indicator (KPI) that measures the amount of revenue generated by the insurance company, per policyholder serviced.
   IMPORTANCE-
   A low, or lagging, value for this KPI could be due to several factors which contribute to below average revenue. We have to look to improve our distribution strategy and investment activities to beef up company earnings.

2. **Top Brokers in Sales Revenue –**
   Measures and ranks the top performing brokers based on sales revenue. By spotting who your top performing brokers are, you can ensure that tentative leads are sent to the experts to be converted.

   IMPORTANCE-
   There are two views to this KPI: one is strategic and takes a look at top sellers over a longer time period (annual), and the other is tactical and looks at short-term performance. Sharing tactical information about this KPI can foster healthy-competition among brokers.

3. **Total premium -** This needs to be done for all personal lines in the agency, not just by carrier.
   Total policies – Also easy to track by totalling all of your policies by carrier into one agency number.
   Total number of clients – This equates to total households- the total number of addresses from your agency management system to get this tally.

   IMPORTANCE-
   These metrics provide us an insight into strength of any agency and how well it can growth for these forecasted even with excessive clients in smaller regions. Also helps us in improving retention plans for such agencies.

<u>Retention/Renewal-It's Sub Metrics:</u>

1. **Renewal/retention –**
   This measures the number of customers who continue coverage after the initial term has expired.
   IMPORTANCE-
   Retention is very cost effective. Many studies have shown that $1 paid towards customer retention increases profits by more than $5 spent on new customer acquisition.

2. **Average length of time clients stay with us –**
   Determining the number of months/years the customer stays with the insurance policy.

IMPORTANCE-

Excellent marketing tactics should deliver a $1.00 return for every $1.00 spent or better in the first year, but you get a much stronger picture for how profitable your marketing is when you know how long you retain your clients on average.

This also helps us keep track of the reasons the customers not retaining the policy enabling us to focus on such grey areas.

METHOD

Basic Steps

1. Loaded the data which was a .csv file by replacing all the ‘99999’ with NA’s.
2. Observe the data and its structure and find out the attributes which made sense and the ones where there is random data .
3. Removal of the unnecessary attributes such as agency_data which also had a lot of NA’s and take only the relevant data.
4. Get the overview if the data by checking the summary and structure of the data.
5. Removed attributes with more than 20% NAs.

Imputation

1. The data which was meant to be imputed was subsetted
2. The subsetted data is now imputed using K-NN imputation.
3. The imputed attributed were PREV_POLY_INFORCE_QTY & PREV_WRTN_PREM_AMT.

SCALING

1. There is an attribute named MONTHS which ranges from 1 to 12.
2. Attributes which are scaled(highlighted in RED) and how it is scales is shown below:

```
for(i in 1:nrow(final_data))


final_data$RETENTION_POLY_QTY[i] <-
final_data$RETENTION_POLY_QTY[i]*(12/final_data$MONTHS[i])

final_data$POLY_INFORCE_QTY[i] <-
final_data$POLY_INFORCE_QTY[i]*(12/final_data$MONTHS[i])
```

```
final_data$PREV_POLY_INFORCE_QTY[i] <-
final_data$PREV_POLY_INFORCE_QTY[i]*(12/final_data$MONTHS[i])

final_data$NB_WRTN_PREM_AMT[i] <-
final_data$NB_WRTN_PREM_AMT[i]*(12/final_data$MONTHS[i])

final_data$WRTN_PREM_AMT[i] <-
final_data$WRTN_PREM_AMT[i]*(12/final_data$MONTHS[i])

final_data$PREV_WRTN_PREM_AMT[i] <-
final_data$PREV_WRTN_PREM_AMT[i]*(12/final_data$MONTHS[i])

final_data$PRD_ERND_PREM_AMT[i] <-
final_data$PRD_ERND_PREM_AMT[i]*(12/final_data$MONTHS[i])

final_data$PRD_INCRD_LOSSES_AMT[i] <-
final_data$PRD_INCRD_LOSSES_AMT[i]*(12/final_data$MONTHS[i])
```

COMPUTATION OF RETENTION RATIO

1. RETENTION_RATIO = (RETENTION_POLY_QTY / PREV_POLY_INFORCE_QTY)
2. When RETENTION_POLY_QTY=0 then

    RETENTION_RATIO=0

   When RETENTION_POLY_QTY !=0 and PREV_POLY_INFORCE_QTY=0 then

    RETENTION_RATIO=MEAN

   Else

    RETENTION_RATIO = (RETENTION_POLY_QTY / PREV_POLY_INFORCE_QTY)

    - Outliers were then detected.
    - The outliers which were detected were detected were substituted by the mean value.

COMPUTATION OF LOSS RATIO

1. LOSS_RATIO =(PRD_INCRD_LOSSES_AMT / WRTN_PREM_AMT)
2. When PRD_INCRD_LOSSES_AMT =0 then

    LOSS_RATIO=0

   When PRD_INCRD_LOSSES_AMT !=0 and WRTN_PREM_AMT =0 then

    LOSS_RATIO=MEAN

   Else

    LOSS_RATIO =(PRD_INCRD_LOSSES_AMT / WRTN_PREM_AMT)

    - Outliers were then detected.
    - The outliers which were detected were detected were substituted by the mean value.

COMPUTATION OF LOSS RATIO(3YRS)

1. Computed by agency by line of business by year for the three year period ending in that year.
2. If there is data for LOSS_RATIO for three years it can be calculated easily by using mean.
3. But for the 1st and 2nd year(i.e 2005 and 2006) the previous two years and the one before 2005 is not available.
4. So to make this more tangible, the first complete year of data we have for an agency will have the loss ratio for that year; the second complete year of data will have the mean loss ratio for that year and the previous year.
5. The third complete year of data will have the mean loss ratio for those three years; then the fourth and greater will have the mean of the three year period ending in that year.

COMPUTATION OF GROWTH RATE

- For the calculation of GROWTH_RATE the following steps are followed

```
for (i in 1:nrow(final_data)
    When (PREV_WRTN_PREM_AMT= 0 && WRTN_PREM_AMT!= 0) then
        GROWTH_RATE =MEAN
    When (PREV_WRTN_PREM_AMT< 0 && WRTN_PREM_AMT > 0) then
        GROWTH_RATE= (WRTN_PREM_AMT - PREV_WRTN_PREM_AMT)
        /-(PREV_WRTN_PREM_AMT)

     When (PREV_WRTN_PREM_AMT= 0)
        GROWTH_RATE = 0

    Else

    GROWTH_RATE =  (WRTN_PREM_AMT - PREV_WRTN_PREM_AMT)/
    PREV_WRTN_PREM_AMT}
```

- Outliers were then detected.
- The outliers which were detected were detected were substituted by the mean value.

COMPUTATION OF GROWTH RATE(3YR)

1. Computed by agency by line of business by year for the three year period ending in that year.
2. If there is data for GROWTH_RATE for three years it can be calculated easily by using mean.
3. But for the 1st and 2nd year(i.e 2005 and 2006) the previous two years and the one before 2005 is not available.
4. So to make this more tangible, the first complete year of data we have for an agency will have the loss ratio for that year; the second complete year of data will have the mean loss ratio for that year and the previous year.

5. The third complete year of data will have the mean loss ratio for those three years; then the fourth and greater will have the mean of the three year period ending in that year.

AGGREGATION

1. Before aggregation we have to remove all the attributes which we used earlier to calculate LOSS_RATIO, RETENTION_RATIO, GROWTH_RATE because those attributes no longer is useful and does not affect the target variables whatsoever.
2. We use the 'dlyr ' package for aggregation.
3. Aggregation (grouping) can be done on basis of
    - AGENCY_ID
    - PROD_ABBR – 33 products of which 14 are CL and 19 are PL
    - PROD_LINE – commercial lines (CL) or personal lines (PL)
    - STATE_ABBR
4. Calculated RETENTION_RATIO, LOSS_RATIO, GROWTH_RATE are found out after aggregation.
5. The above mentioned attributes will be responsible to predict the agency performance.
6. After grouping computed the mean of the target attributes

BINNING

1. Calculated RETENTION_RATIO, LOSS_RATIO, GROWTH_RATE are binned appropriately(high or low)

- Decision trees are the applied by keeping RETENTION_RATIO, LOSS_RATIO or GROWTH_RATE as the targer variables

DATA

Explanation of Agency Data Set

- AGENCY_ID
- PRIMARY_AGENCY_ID – (contains missing values) master agency if part of group
- PROD_ABBR – 33 products of which 14 are CL and 19 are PL
- PROD_LINE – commercial lines (CL) or personal lines (PL)
- STATE_ABBR
- STAT_PROFILE_DATE_YEAR – data starts in mid-2005 and continues into 2015
- RETENTION_POLY_QTY – current number of policies that are still active from previous year
- POLY_INFORCE_QTY – number of policies active for that year
- PREV_POLY_INFORCE_QTY – (contains missing values) number of policies active in the previous year
- NB_WRTN_PREM_AMT – new business in written premium
- WRTN_PREM_AMT – total written premium
- PREV_WRTN_PREM_AMT – (contains missing values) written premium during the same period in the previous year
- PRD_ERND_PREM_AMT – amount of premium taken in
- PRD_INCRD_LOSSES_AMT – losses
- MONTHS – number of months included in the data for that year; the original data was monthly and some months were missing so the aggregate doesn't make up the entire year if the months value is less than 12
- RETENTION_RATIO – (computed & contains missing values) computed for each row in the data as RETENTION_POLY_QTY / PREV_POLY_INFORCE_QTY; therefore it's a granular measure of the retention for that agency writing that particular product in that particular state from the previous year
- LOSS_RATIO – (computed & contains missing values) computed for each row in the data as PRD_INCRD_LOSSES_AMT / WRTN_PREM_AMT; currently I'm only computing results where the WRTN_PREM_AMT is greater than 0; however there are many cases where there are losses

but no premium maybe from a previous claim that's still being paid, so I included codes to indicate whether there are positive or negative losses on zero premiums; they are located at the end of this document

- LOSS_RATIO_3YR – (computed & contains missing values) computed by agency by line of business by year for the three year period ending in that year, if there is data for three years, otherwise the two years or one year of data available; to make this more tangible, the first complete year of data we have for an agency will have the loss ratio for that year; the second complete year of data will have the mean loss ratio for that year and the previous year; the third complete year of data will have the mean loss ratio for those three years; then the fourth and greater will have the mean of the three year period ending in that year; note that the mean loss ratios are computed independently for PL and CL
- GROWTH_RATE_3YR – (computed & contains missing values) computed by agency by line of business by year for the three year period ending in that year; measures the average growth in written premium for that agency in that line of business; only computes results for agencies that have data for the entire range of years; since the measure is over three years of growth, there needs to be a base year to be used as a standard so four years of data are needed; in order to include as many results as possible, the PREV_WRTN_PREM_AMT column is used if it exists in the first year of data available, so that it can be used as a base year, otherwise the WRTN_PREM_AMT is the only column used
- AGENCY_APPOINTMENT_YEAR – (contains missing values) year the agency started doing business with Azure
- ACTIVE_PRODUCERS – (contains missing values) number of active producers in the agency
- MAX_AGE – (contains missing values & results may not be accurate) maximum age producer at that agency
- MIN_AGE – (contains missing values) minimum age producer at that agency
- VENDOR_IND – indicator column to specify whether the agency subscribes to a vendor
- VENDOR – (contains missing values) the vendor that the agency subscribes to
- PL_START_YEAR – (contains missing values) year the agency started using the PL vendor
- PL_END_YEAR – (contains missing values) year the agency stopped using the PL vendor

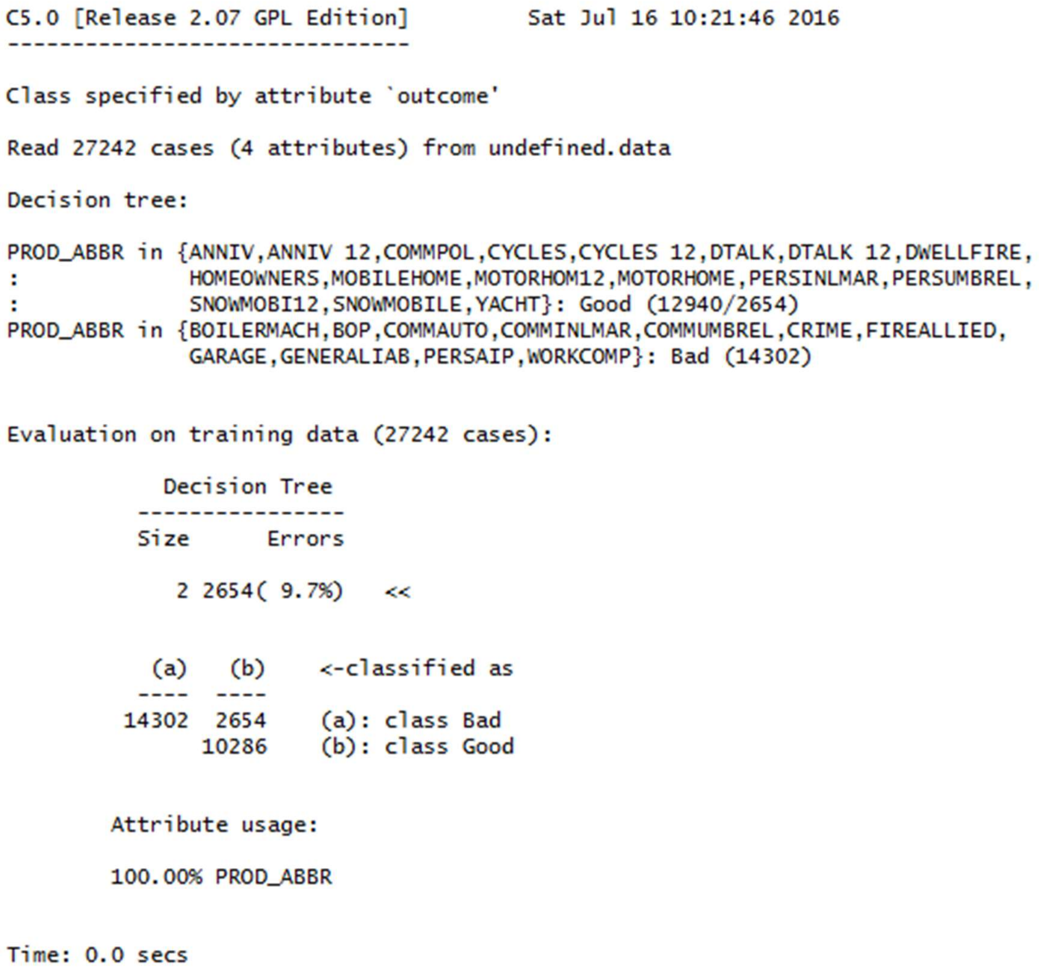
- COMMISIONS_START_YEAR – (contains missing values) year the agency started using the COMMISIONS vendor
- COMMISIONS_END_YEAR – (contains missing values) year the agency stopped using the COMMISIONS vendor
- CL_START_YEAR – (contains missing values) year the agency started using the CL vendor
- CL_END_YEAR – (contains missing values) year the agency stopped using the CL vendor
- ACTIVITY_NOTES_START_YEAR – (contains missing values) year the agency started using the ACTIVITY NOTES vendor
- ACTIVITY_NOTES_END_YEAR – (contains missing values) year the agency stopped using the ACTIVITY NOTES vendor
- CL_BOUND_CT_MDS – (contains missing values) number of bound policies quoted through a MDS (probably a data recording error, should be DSM) in the current year to date, that is the first six months of 2015, in commercial lines
- CL_QUO_CT_MDS – (contains missing values) number of quoted policies through a MDS (probably a data recording error, should be DSM) in the current year to date, that is the first six months of 2015, in commercial lines
- CL_BOUND_CT_SBZ – (contains missing values) number of bound policies quoted through a SBZ in the current year to date, that is the first six months of 2015, in commercial lines
- CL_QUO_CT_SBZ – (contains missing values) number of quoted policies through a SBZ in the current year to date, that is the first six months of 2015, in commercial lines
- CL_BOUND_CT_eQT – (contains missing values) number of bound policies quoted through an eQT in the current year to date, that is the first six months of 2015, in commercial lines
- CL_QUO_CT_eQT – (contains missing values) number of quoted policies though an eQT in the current year to date, that is the first six months of 2015, in commercial lines
- PL_BOUND_CT_ELINKS – (contains missing values) number of bound policies quoted through ELINKS since September 2013 in personal lines
- PL_QUO_CT_ELINKS – (contains missing values) number of quoted policies though ELINKS since September 2013 in personal lines
- PL_BOUND_CT_PLRANK – (contains missing values) number of bound policies quoted through PLRANK since September 2013 in personal lines
- PL_QUO_CT_PLRANK – (contains missing values) number of quoted policies though PLRANK since September 2013 in personal lines

- PL_BOUND_CT_eQTte – (contains missing values) number of bound policies quoted through eQTte since September 2013 in personal lines
- PL_QUO_CT_eQTte – (contains missing values) number of quoted policies though eQTte since September 2013 in personal lines

RESULTS

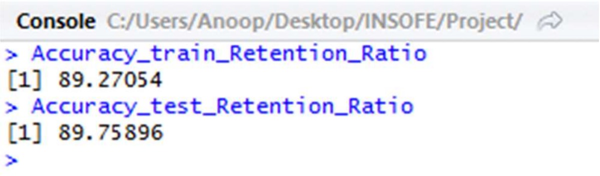
OUTPUT:

RETENTION_RATIO AS PERFORMANCE INDICATOR:



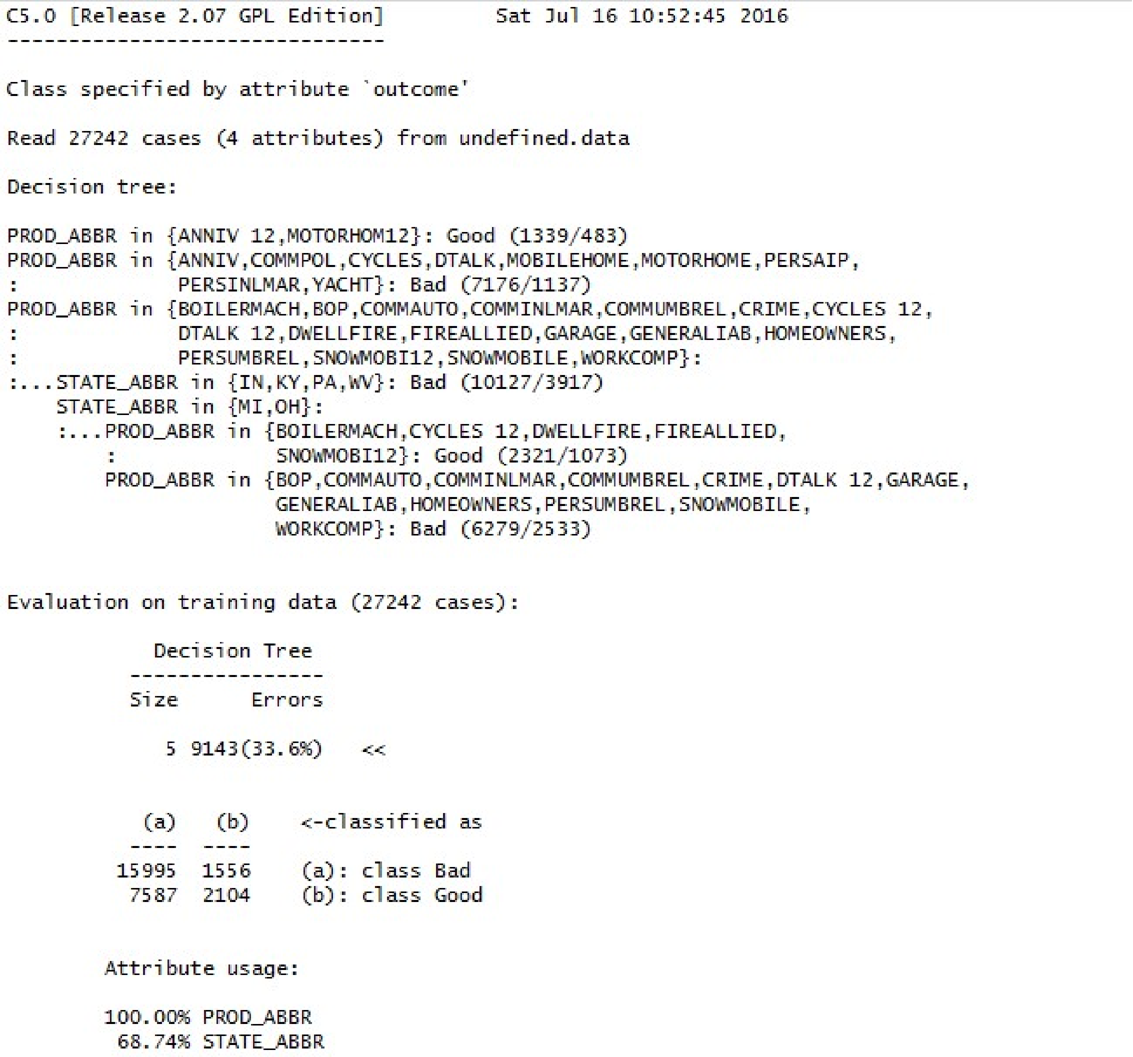
```
C5.0 [Release 2.07 GPL Edition]        Sat Jul 16 10:21:46 2016
-------------------------------

Class specified by attribute `outcome'

Read 27242 cases (4 attributes) from undefined.data

Decision tree:

PROD_ABBR in {ANNIV,ANNIV 12,COMMPOL,CYCLES,CYCLES 12,DTALK,DTALK 12,DWELLFIRE,
:              HOMEOWNERS,MOBILEHOME,MOTORHOM12,MOTORHOME,PERSINLMAR,PERSUMBREL,
:              SNOWMOBI12,SNOWMOBILE,YACHT}: Good (12940/2654)
PROD_ABBR in {BOILERMACH,BOP,COMMAUTO,COMMINLMAR,COMMUMBREL,CRIME,FIREALLIED,
              GARAGE,GENERALIAB,PERSAIP,WORKCOMP}: Bad (14302)


Evaluation on training data (27242 cases):

            Decision Tree
          ----------------
          Size      Errors

             2 2654( 9.7%)   <<


           (a)   (b)    <-classified as
          ----  ----
         14302  2654    (a): class Bad
               10286    (b): class Good


        Attribute usage:

        100.00% PROD_ABBR


Time: 0.0 secs
```

ACCURACY:



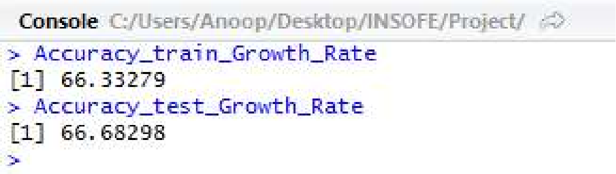
```
> Accuracy_train_Retention_Ratio
[1] 89.27054
> Accuracy_test_Retention_Ratio
[1] 89.75896
>
```

GROWTH_RATE AS PERFORMANCE INDICATOR:



```
C5.0 [Release 2.07 GPL Edition]        Sat Jul 16 10:52:45 2016
-------------------------------

Class specified by attribute `outcome'

Read 27242 cases (4 attributes) from undefined.data

Decision tree:

PROD_ABBR in {ANNIV 12,MOTORHOM12}: Good (1339/483)
PROD_ABBR in {ANNIV,COMMPOL,CYCLES,DTALK,MOBILEHOME,MOTORHOME,PERSAIP,
:             PERSINLMAR,YACHT}: Bad (7176/1137)
PROD_ABBR in {BOILERMACH,BOP,COMMAUTO,COMMINLMAR,COMMUMBREL,CRIME,CYCLES 12,
:             DTALK 12,DWELLFIRE,FIREALLIED,GARAGE,GENERALIAB,HOMEOWNERS,
:             PERSUMBREL,SNOWMOBI12,SNOWMOBILE,WORKCOMP}:
:...STATE_ABBR in {IN,KY,PA,WV}: Bad (10127/3917)
    STATE_ABBR in {MI,OH}:
    :...PROD_ABBR in {BOILERMACH,CYCLES 12,DWELLFIRE,FIREALLIED,
        :             SNOWMOBI12}: Good (2321/1073)
        PROD_ABBR in {BOP,COMMAUTO,COMMINLMAR,COMMUMBREL,CRIME,DTALK 12,GARAGE,
                      GENERALIAB,HOMEOWNERS,PERSUMBREL,SNOWMOBILE,
                      WORKCOMP}: Bad (6279/2533)


Evaluation on training data (27242 cases):

	    Decision Tree
	  ----------------
	  Size      Errors

	     5 9143(33.6%)   <<


	    (a)   (b)    <-classified as
	   ----  ----
	  15995  1556    (a): class Bad
	   7587  2104    (b): class Good


	Attribute usage:

	100.00% PROD_ABBR
	 68.74% STATE_ABBR
```

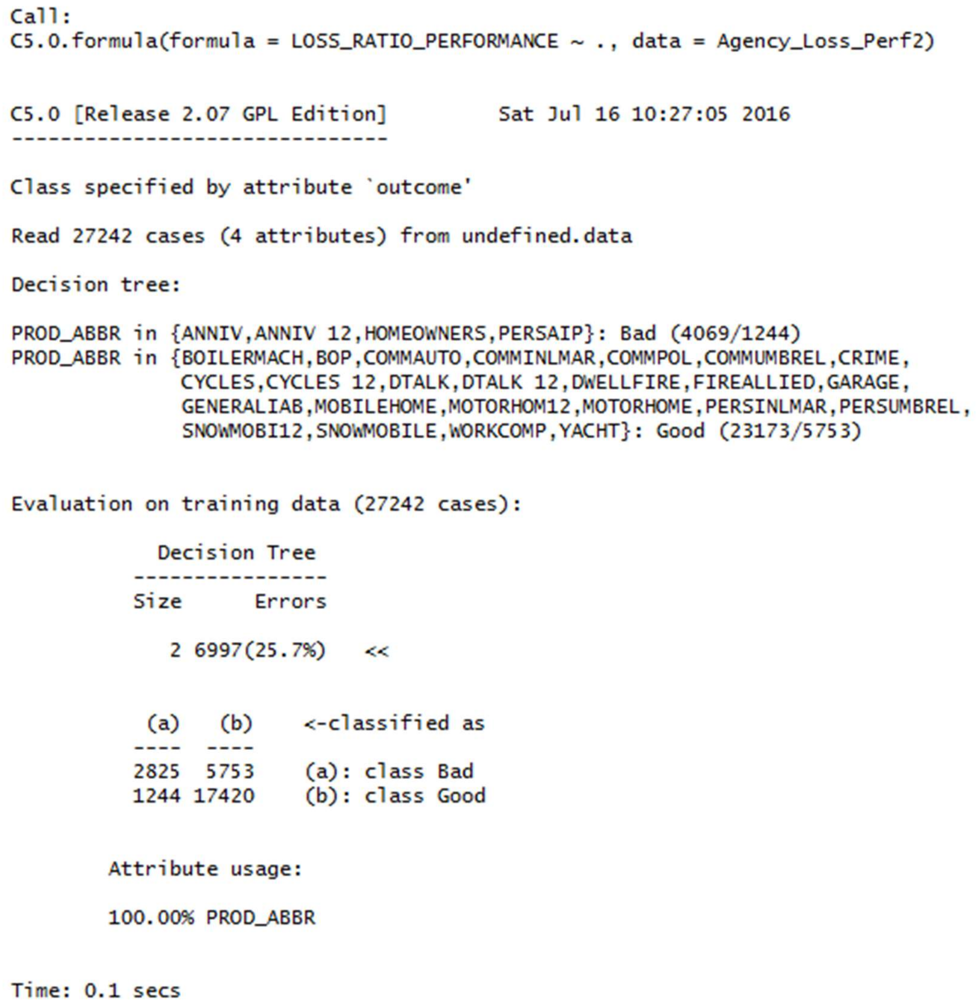
ACCURACY:



Console C:/Users/Anoop/Desktop/INSOFE/Project/

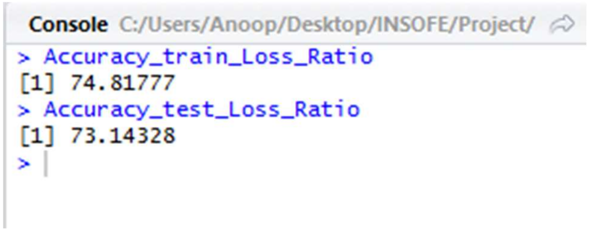
```
> Accuracy_train_Growth_Rate
[1] 66.33279
> Accuracy_test_Growth_Rate
[1] 66.68298
>
```

LOSS_RATIO AS PERFORMANCE INDICATOR:



```
Call:
C5.0.formula(formula = LOSS_RATIO_PERFORMANCE ~ ., data = Agency_Loss_Perf2)


C5.0 [Release 2.07 GPL Edition]        Sat Jul 16 10:27:05 2016
-------------------------------

Class specified by attribute `outcome'

Read 27242 cases (4 attributes) from undefined.data

Decision tree:

PROD_ABBR in {ANNIV,ANNIV 12,HOMEOWNERS,PERSAIP}: Bad (4069/1244)
PROD_ABBR in {BOILERMACH,BOP,COMMAUTO,COMMINLMAR,COMMPOL,COMMUMBREL,CRIME,
              CYCLES,CYCLES 12,DTALK,DTALK 12,DWELLFIRE,FIREALLIED,GARAGE,
              GENERALIAB,MOBILEHOME,MOTORHOM12,MOTORHOME,PERSINLMAR,PERSUMBREL,
              SNOWMOBI12,SNOWMOBILE,WORKCOMP,YACHT}: Good (23173/5753)


Evaluation on training data (27242 cases):

	    Decision Tree
	  ----------------
	  Size      Errors

	     2 6997(25.7%)   <<


	   (a)   (b)    <-classified as
	  ----  ----
	  2825  5753    (a): class Bad
	  1244 17420    (b): class Good


	Attribute usage:

	100.00% PROD_ABBR


Time: 0.1 secs
```

ACCURACY:



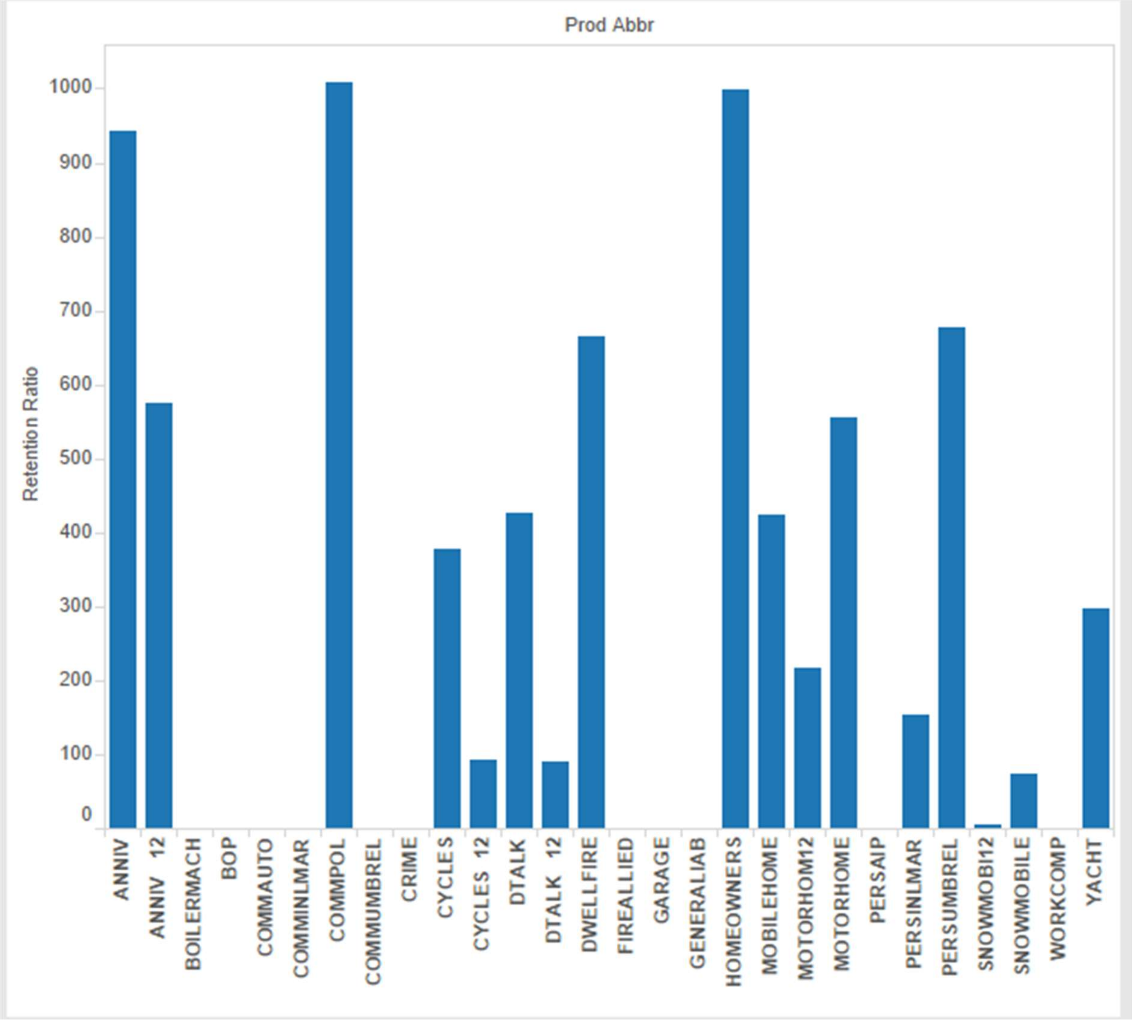
```
> Accuracy_train_Loss_Ratio
[1] 74.81777
> Accuracy_test_Loss_Ratio
[1] 73.14328
> 
```

ANALYSIS

RETENTION RATIO:

This graph clearly supports the rules that is generated when RETENTION RATIO is kept as the target variable.

The products which has very low RETENTION RATIO can be evidently seen in the bar graph.



<u>GROWTH RATE</u>:

This graph clearly supports the rules that is generated when GROWTH RATE is kept as the target variable.

Wherever there is bad GROWTH RATE there is a very low bar and vice-versa for good GROWTH RATE.

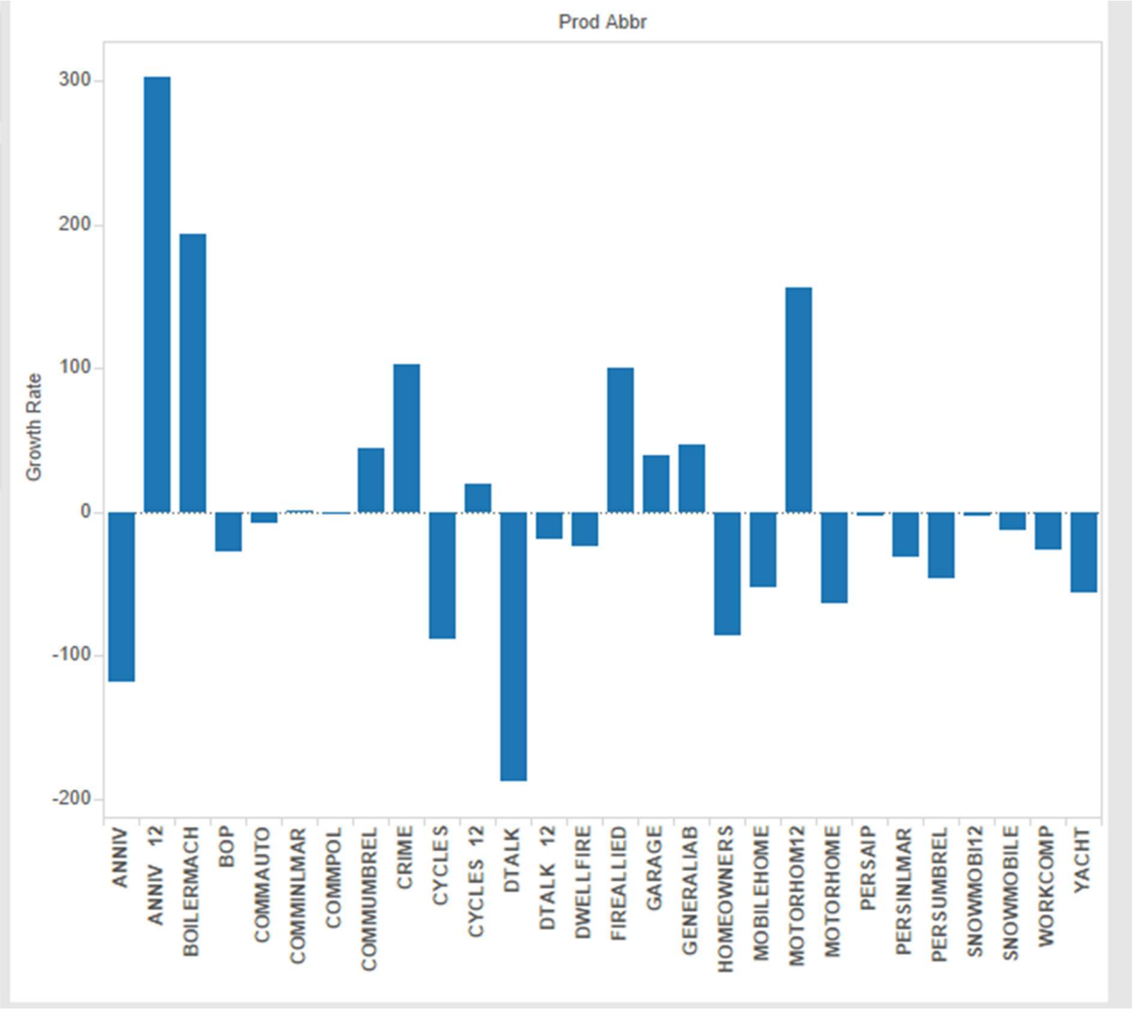


Fig 1

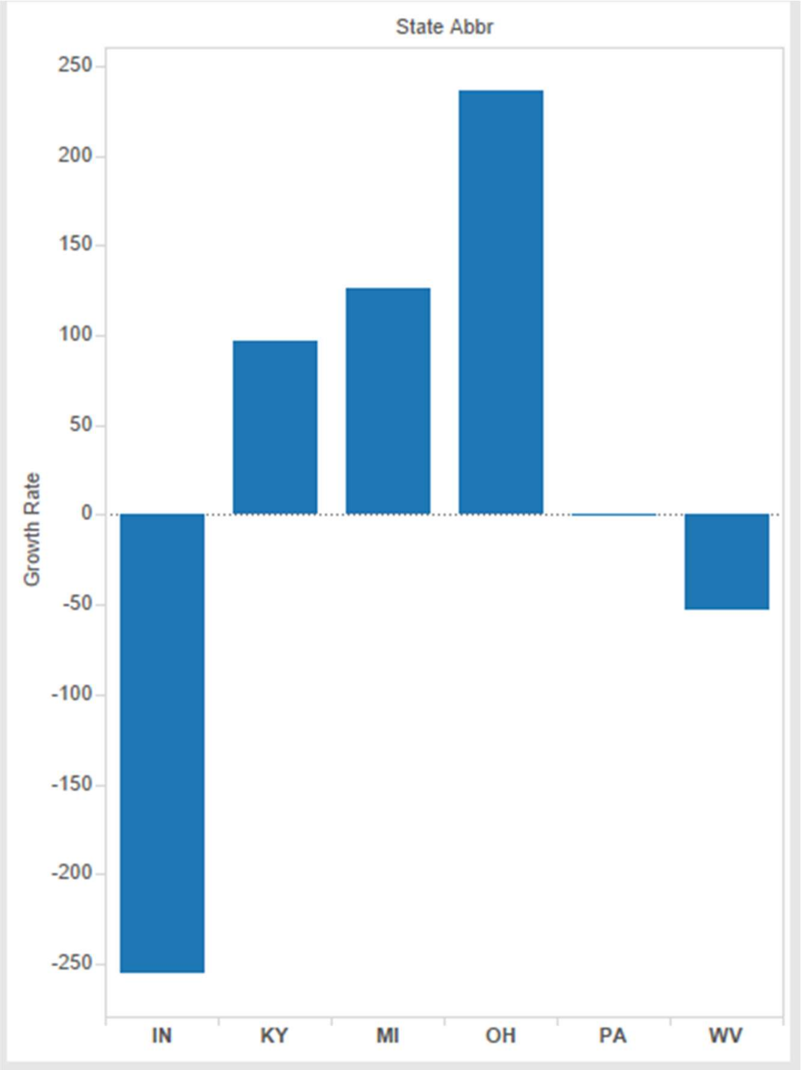

State Abbr
Growth Rate
250
200
150
100
50
0
-50
-100
-150
-200
-250
IN
KY
MI
OH
PA
WV

Fig 2

LOSS RATIO:

ANNIV has the most LOSS RATIO as shown the consecutive figures and the bar graphs below explain the rules that were generated.

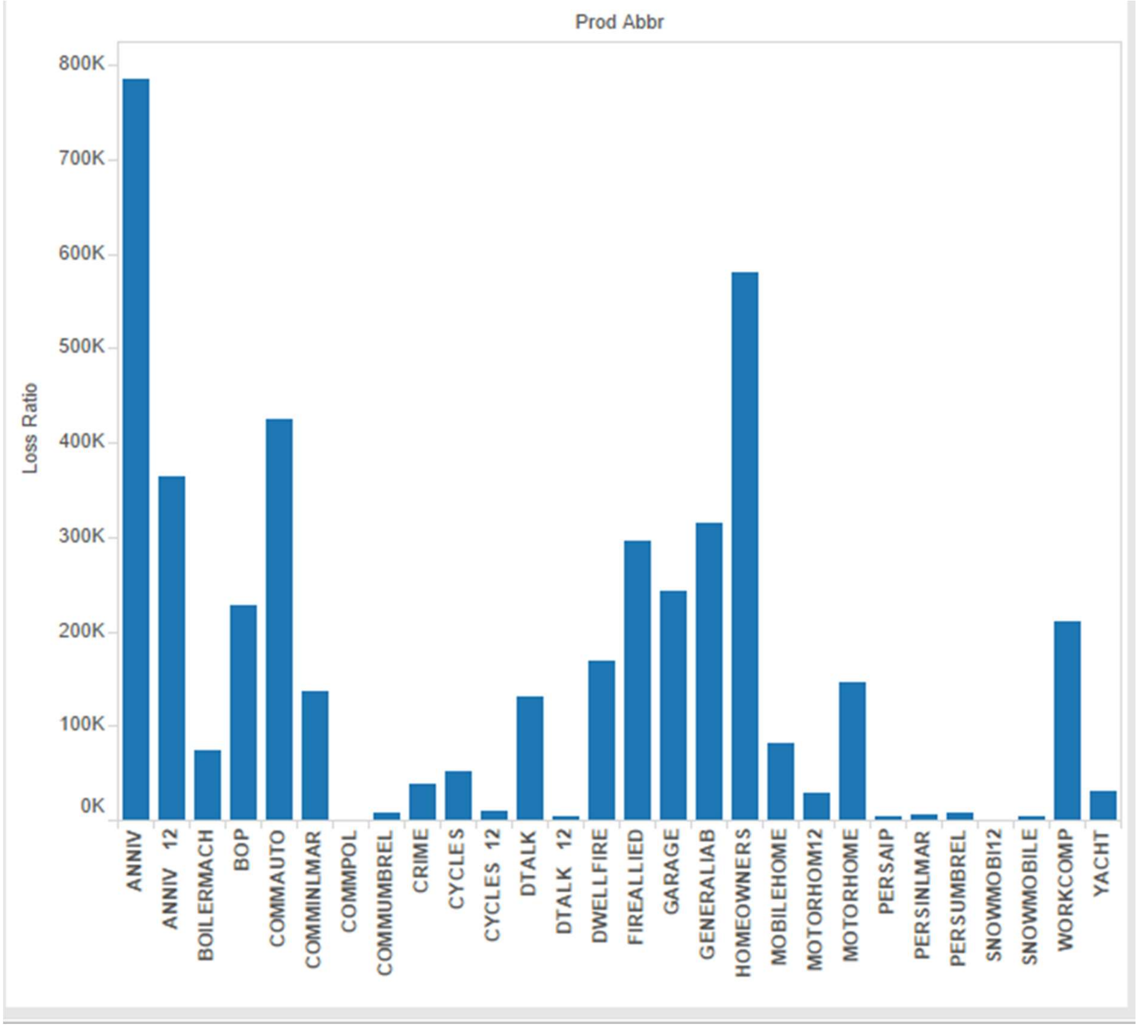


Fig 1

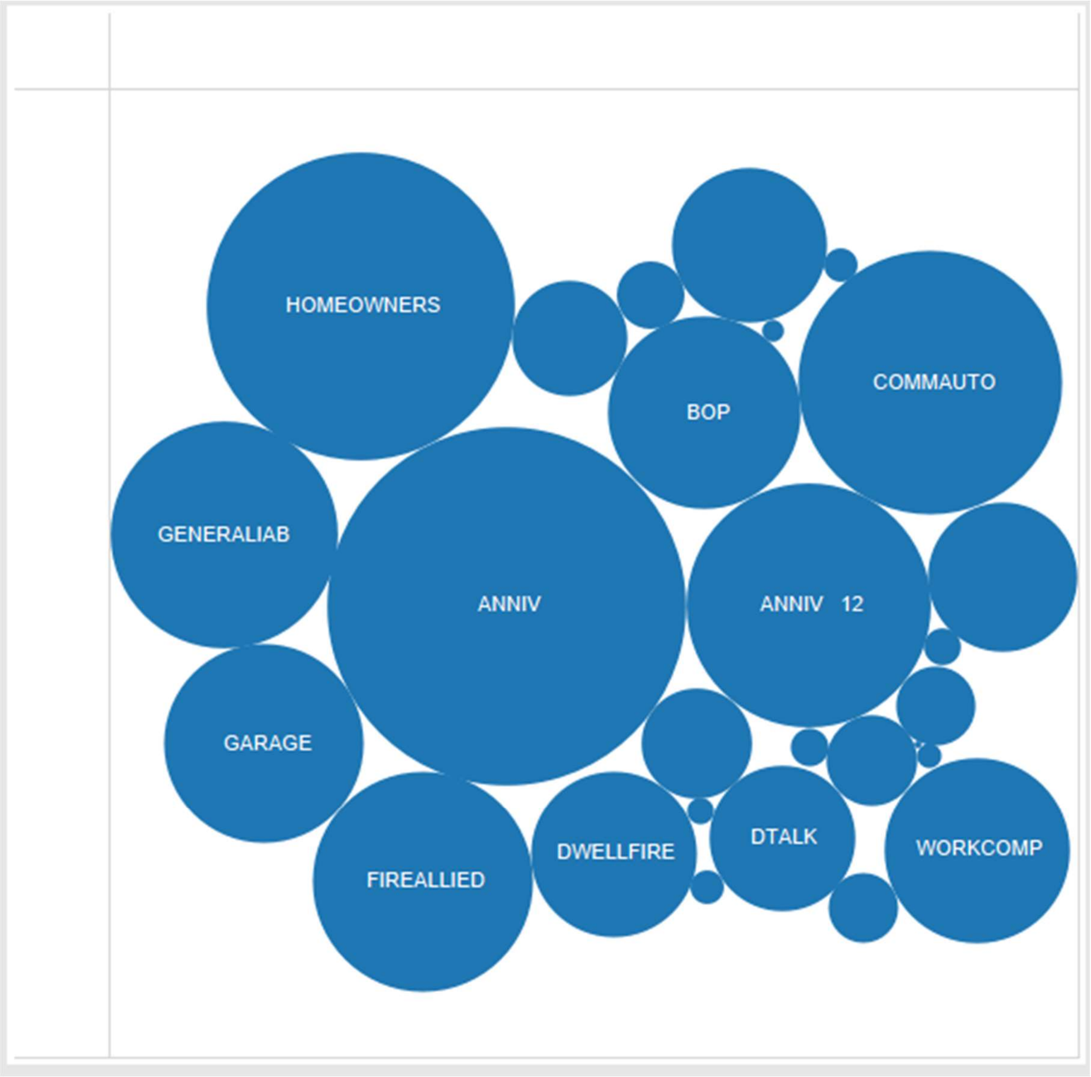

HOMEOWNERS
BOP
COMMAUTO
GENERALIAB
ANNIV
ANNIV 12
GARAGE
FIREALLIED
DWELLFIRE
DTALK
WORKCOMP

Fig 2

APPENDICES

Anoop_Madamsetty_PROJECT.R