Project Report

On

Collaborative Filtering for Cross Selling of Insurance

By

Saurabh Manjrekar

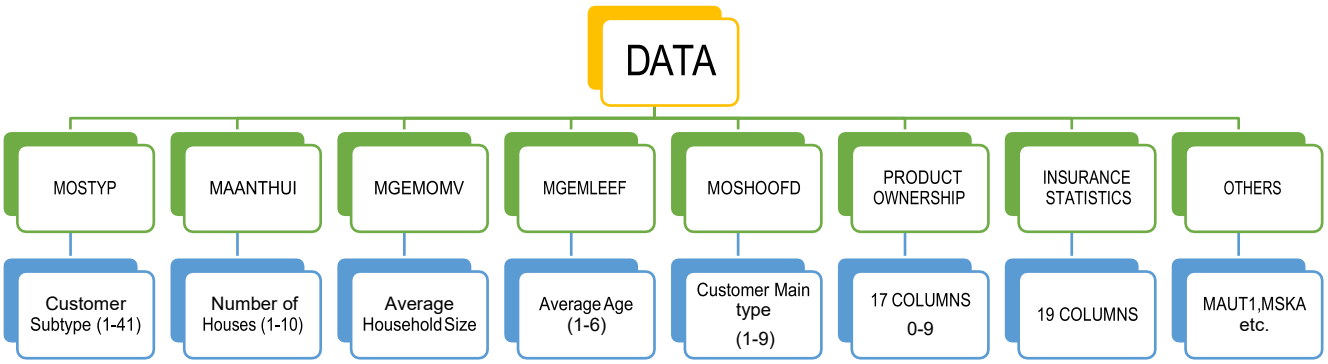
Table of Contents

1.0 Introduction

Insurance in the United States refers to the market for risk, the world's largest insurance market by premium volume. Of the $4.640 trillion of gross premiums written worldwide in 2013, $1.274 trillion (27%) were written in the United States.

In our project we are helping the insurance firms to identify the cross-selling opportunity for Insurance products based on the Demographic, Physiological, Socio-Economic and Policy ownership profile of the customers by devising an automated Collaborative filtering engine. This will help them in retaining their existing customers and also will decrease their customer acquisition cost and hence increase their revenue.

2.0 Dataset



Data Source: Kaggle

No of observations: 9822 real customer records

No of variables: 87

- Each real customer record consists of 87 variables, containing sociodemographic data (variables 1-44) and product ownership data (variables 45-87).
    - The sociodemographic data is derived from zip codes. All customers living in areas with the same zip code have the same sociodemographic attributes.
    - Variable 1- 44 are socio-demographic data and variables 45-87 are product ownership data.
- The socio-demographic data is given by zip code. All customers living in the same zip code have the same sociodemographic attributes.
- Variables beginning with M refer to demographic statistics of the postal code, while variables beginning with P and A (as well as CARAVAN) refer to product ownership and insurance statistics in the postal code.

3.0 Feature Selection

We had 87 features in our dataset of which not all were of important so we did feature manipulation to fetch the relevant features from the dataset to help the model perform better and more efficiently. We finally extracted 54 features that were relevant for the model building.

Total number of independent variables remaining before applying the techniques: 87

The two approaches used to extract the relevant features are:

3.1 Correlation Matrix

Correlation states how the features are related to each other. Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. So, when two features have high correlation, we can drop one of the two features. Correlation matrix or Heat-Map makes it easy to identify which features are most related to each other.
We plotted the correlation matrix to check multi-collinearity amongst the independent variables and removed one of the features that were having correlation higher than 0.7
The features which were removed after applying correlation are:

'ORIGIN', 'MOSHOOFD', 'MGODGE', 'MRELOV', 'MFALLEEN', 'MFWEKIND', 'MOPLHOOG', 'MOPLMIDD', 'MBERHOOG', 'MHKOOP', 'MAUT1', 'MZFONDS', 'PWAPART', 'PWABEDR', 'PWALAND', 'PPERSAUT', 'PBESAUT', 'PMOTSCO', 'PVRAAUT', 'PAANHANG', 'PTRACTOR', 'PWERKT', 'PBROM', 'PLEVEN', 'PPERSONG', 'PGEZONG', 'PWAOREG', 'PBRAND', 'PZEILPL', 'PPLEZIER', 'PFIETS', 'PINBOED', 'PBYSTAND'

Total number of independent variables remaining after applying this technique: 54

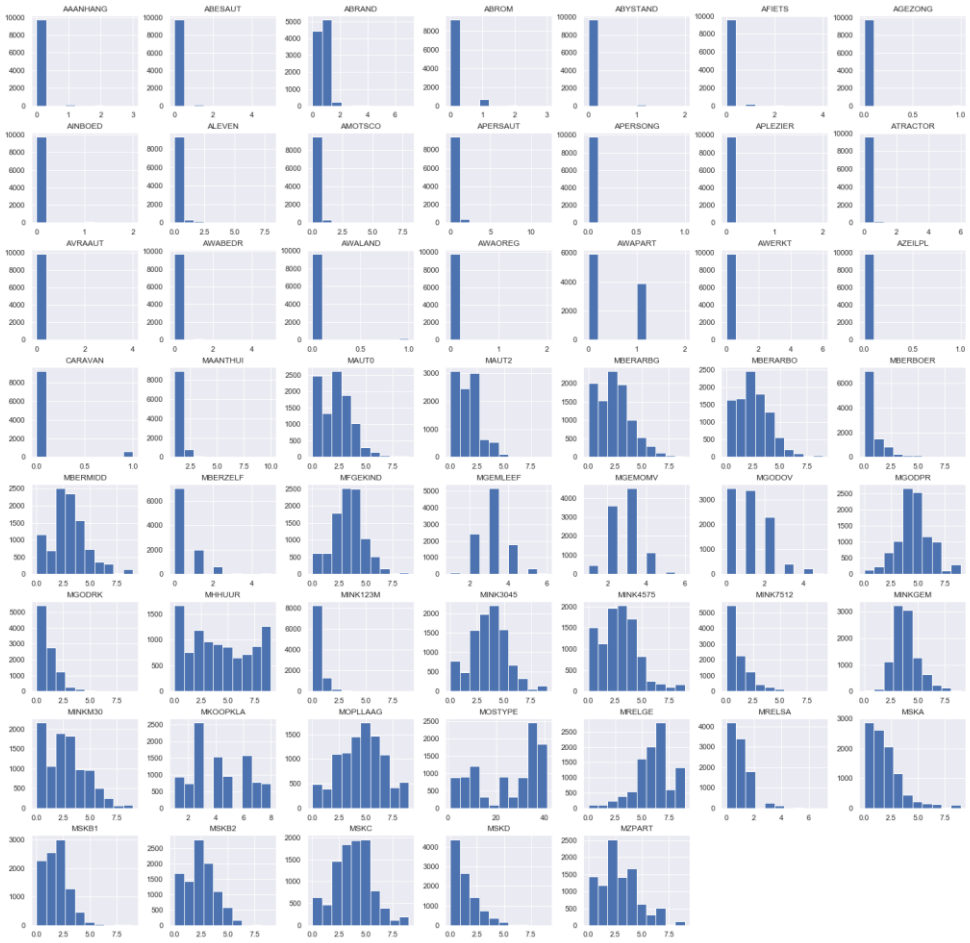
```
In [101]: # dataset.iloc[:,1:-1].corr()
          import seaborn as sns # data visualization library
          import matplotlib.pyplot as plt
          #correlation map
          f,ax = plt.subplots(figsize=(50, 50))
          sns.heatmap(dataset.iloc[:,1:-1].corr(), annot=True, linewidths=.5, fmt= '.1f',ax=ax)
Out[101]: <matplotlib.axes._subplots.AxesSubplot at 0x1f66a15cb00>
```

3.2 Improper Distribution

This technique was used to verify variables which are imbalanced.

We plotted histograms for the remaining variables (except product variables beginning with P, A as well as CARAVAN) and we did not find any imbalance data

Total number of independent variables remaining after applying this technique: 54



4.0 Model Building

For this Dataset, we have implemented User Collaborative Recommendation Engine. Here, we try to search for lookalike customers and offer products based on what his/her lookalike has chosen. This algorithm is very effective but takes a lot of time and resources. For example, if a person A has taken insurances 1, 2, 3 and B has taken insurances 1,2,3,4 then they have similar interests and A should be recommended 4th insurance.

Extracting the list of policies which current user doesn't have

[112]:
```
availablecollist=[]
for item in policiescolumnlist:
    if df.loc[user_index,item]==0:
        availablecollist.append(item)
```

Model for User-Collaborative Filtering

[127]:
```
no_of_policies=4 #you can set this variable value as number of policies you want to recommend.
output=dict()
i=1
# print(output)
while i < 50:
#     print("OK")
    opdf=df.iloc[user_similarity[user_index].argsort()[(i-1)*10:i*10], 31:53] #list of indexes of row
#     print(opdf[availablecollist].sum().sort_values()[opdf[availablecollist].sum().sort_values()>0])
    tempoutput=dict(opdf[availablecollist].sum().sort_values()[opdf[availablecollist].sum().sort_values()>0])
    output.update(tempoutput)
#     print(len(output))
#     print(output)
#     if len(output)>=no_of_policies:
#         break
    i+=1
    if i>50:
        print("There are no products, as per top 50 similar people")
        break
```

Output for User in row 9:

In [140]:
```
output
```

Out[140]:
```
{'CARAVAN': 1, 'AAANHANG': 1, 'ABESAUT': 3, 'AWAPART': 5}
```

In [141]:
```
print(sorted(output))
```

```
['AAANHANG', 'ABESAUT', 'AWAPART', 'CARAVAN']
```

Here, user-9 has been suggested to purchase 4 insurance products (AAANHANG, ABESAUT, AWAPART, CARAVAN). It also gives the count of customer who are alike with user-9 and has the respective insurance product.
For example, ABESAUT: 3 states that there are 3 other users like user-9 and they have ABESAUT as well.
Similarly, a list of recommended products will be given to users among which he can purchase the next insurance product.

5.0 Conclusion

In our project we started with 87 variables which contained the details about the demo-graphic, socio-economic and policy details of the all the existing customers. With the help of feature engineering we reduced these 87 variables into 54 variables which helped us in modeling a recommender engine to suggest customer about which insurance to be purchased.

6.0 Appendix

- ORIGIN: *train* or *test*, as described above
- MOSTYPE: Customer Subtype; see L0
- MAANTHUI: Number of houses (1 – 10)
- MGEMOMV: Avg size household (1 – 6)
- MGEMLEEF: Avg age; ***see L1***
- MOSHOOFD: Customer main type; ***see L2***

*** Percentages in each group, per postal code (**see L3**)**:*

- MGODRK: Roman catholic
- MGODPR: Protestant ...
- MGODOV: Other religion
- MGODGE: No religion
- MRELGE: Married
- MRELSA: Living together
- MRELOV: Other relation
- MFALLEEN: Singles
- MFGEKIND: Household without children
- MFWEKIND: Household with children
- MOPLHOOG: High level education
- MOPLMIDD: Medium level education
- MOPLLAAG: Lower level education
- MBERHOOG: High status
- MBERZELF: Entrepreneur
- MBERBOER: Farmer
- MBERMIDD: Middle management
- MBERARBG: Skilled labourers
- MBERARBO: Unskilled labourers
- MSKA: Social class A
- MSKB1: Social class B1
- MSKB2: Social class B2
- MSKC: Social class C
- MSKD: Social class D
- MHHUUR: Rented house
- MHKOOP: Home owners
- MAUT1: 1 car
- MAUT2: 2 cars
- MAUT0: No car
- MZFONDS: National Health Service
- MZPART: Private health insurance
- MINKM30: Income < 30.000
- MINK3045: Income 30-45.000
- MINK4575: Income 45-75.000
- MINK7512: Income 75-122.000
- MINK123M: Income >123.000
- MINKGEM: Average income

- MKOOPKLA: Purchasing power class

** Total number of variable in postal code (***see L4***) **:

- PWAPART: Contribution private third party insurance
- PWABEDR: Contribution third party insurance (firms) ...
- PWALAND: Contribution third party insurane (agriculture)
- PPERSAUT: Contribution car policies
- PBESAUT: Contribution delivery van policies
- PMOTSCO: Contribution motorcycle/scooter policies
- PVRAAUT: Contribution lorry policies
- PAANHANG: Contribution trailer policies
- PTRACTOR: Contribution tractor policies
- PWERKT: Contribution agricultural machines policies
- PBROM: Contribution moped policies
- PLEVEN: Contribution life insurances
- PPERSONG: Contribution private accident insurance policies
- PGEZONG: Contribution family accidents insurance policies
- PWAOREG: Contribution disability insurance policies
- PBRAND: Contribution fire policies
- PZEILPL: Contribution surfboard policies
- PPLEZIER: Contribution boat policies
- PFIETS: Contribution bicycle policies
- PINBOED: Contribution property insurance policies
- PBYSTAND: Contribution social security insurance policies
- AWAPART: Number of private third party insurance 1 - 12
- AWABEDR: Number of third party insurance (firms) ...
- AWALAND: Number of third party insurance (agriculture)
- APERSAUT: Number of car policies
- ABESAUT: Number of delivery van policies
- AMOTSCO: Number of motorcycle/scooter policies
- AVRAAUT: Number of lorry policies
- AAANHANG: Number of trailer policies
- ATRACTOR: Number of tractor policies
- AWERKT: Number of agricultural machines policies
- ABROM: Number of moped policies
- ALEVEN: Number of life insurances
- APERSONG: Number of private accident insurance policies
- AGEZONG: Number of family accidents insurance policies
- AWAOREG: Number of disability insurance policies
- ABRAND: Number of fire policies
- AZEILPL: Number of surfboard policies
- APLEZIER: Number of boat policies
- AFIETS: Number of bicycle policies
- AINBOED: Number of property insurance policies
- ABYSTAND: Number of social security insurance policies
- CARAVAN: Number of mobile home policies 0 – 1

All data we have is already encoded in following way:

L0: Customer subtype

1: High Income, expensive child
2: Very Important Provincials
3: High status seniors
4: Affluent senior apartments
5: Mixed seniors

6: Career and childcare
7: Dinky's (double income no kids)
8: Middle class families
9: Modern, complete families
10: Stable family
11: Family starters
12: Affluent young families
13: Young all American family

14: Junior cosmopolitan
15: Senior cosmopolitans

16: Students in apartments
17: Fresh masters in the city
18: Single youth
19: Suburban youth
20: Ethnically diverse
21: Young urban have-nots
22: Mixed apartment dwellers

23: Young and rising
24: Young, low educated
25: Young seniors in the city
26: Own home elderly
27: Seniors in apartments
28: Residential elderly
29: Porch less seniors: no front yard
30: Religious elderly singles
31: Low income Catholics

32: Mixed seniors
33: Lower class large families
34: Large family, employed child
35: Village families
36: Couples with teens 'Married with children'
37: Mixed small town dwellers
38: Traditional families
39: Large religious families
40: Large family farms
41: Mixed rural

L1: average age keys:

1: 20-30 years
2: 30-40 years
3: 40-50 years
4: 50-60 years
5: 60-70 years
6: 70-80 years

L2: customer main type keys:

1: Successful hedonists
2: Driven Growers
3: Average Family
4: Career Loners
5: Living well
6: Cruising Seniors
7: Retired and Religious
8: Family with grown ups
9: Conservative families 10: Farmers

L3: percentage keys:

0: 0%
1: 1 - 10%
2: 11 - 23%
3: 24 - 36%
4: 37 - 49%
5: 50 - 62%
6: 63 - 75%
7: 76 - 88%
8: 89 - 99%
9: 100%

L4: total number keys:

0: 0
1: 1 - 49
2: 50 - 99
3: 100 - 199
4: 200 - 499
5: 500 - 999
6: 1000 - 4999
7: 5000 - 9999
8: 10,000 - 19,999
9: >= 20,000