Sales Optimization Proposal

Minh Phuong

TOC

Overview

Company A has faced declining sales for the past half year due to competition. Its Sales Managers have proposed to implement data science and modelling to look for opportunities to increase sales.

Company A also believes that the customer's past purchases would affect his/her decision in purchasing the next product. They have requested to find a way to take into consideration what each customer has prior.

Problems to solve

1. Exploratory Data Analysis
2. Descriptive Statistics Model
3. Clustering
4. Recommendation Engine

Project objective

Increase up-selling and cross-selling efforts using Data Science

Understanding the data

Data Exploration

00

Some characteristics of the dataset can show certain essential consensus information such as average weight, height, age, etc...

```
No of numerical:  12
No of categorical:  37
```

COVER_TERM	PAY_TERM	PAY_AGE	COVER_AGE	ENTRY_AGE
20752.000000	20752.000000	20752.000000	20752.000000	20752.000000
5.938657	5.938657	35.492772	35.492772	30.918658
13.478252	13.478252	22.805384	22.805384	18.898141
0.000000	0.000000	0.000000	0.000000	0.000000
1.000000	1.000000	16.000000	16.000000	13.000000
1.000000	1.000000	36.000000	36.000000	32.000000
5.000000	5.000000	52.000000	52.000000	46.000000
100.000000	100.000000	109.000000	109.000000	75.000000

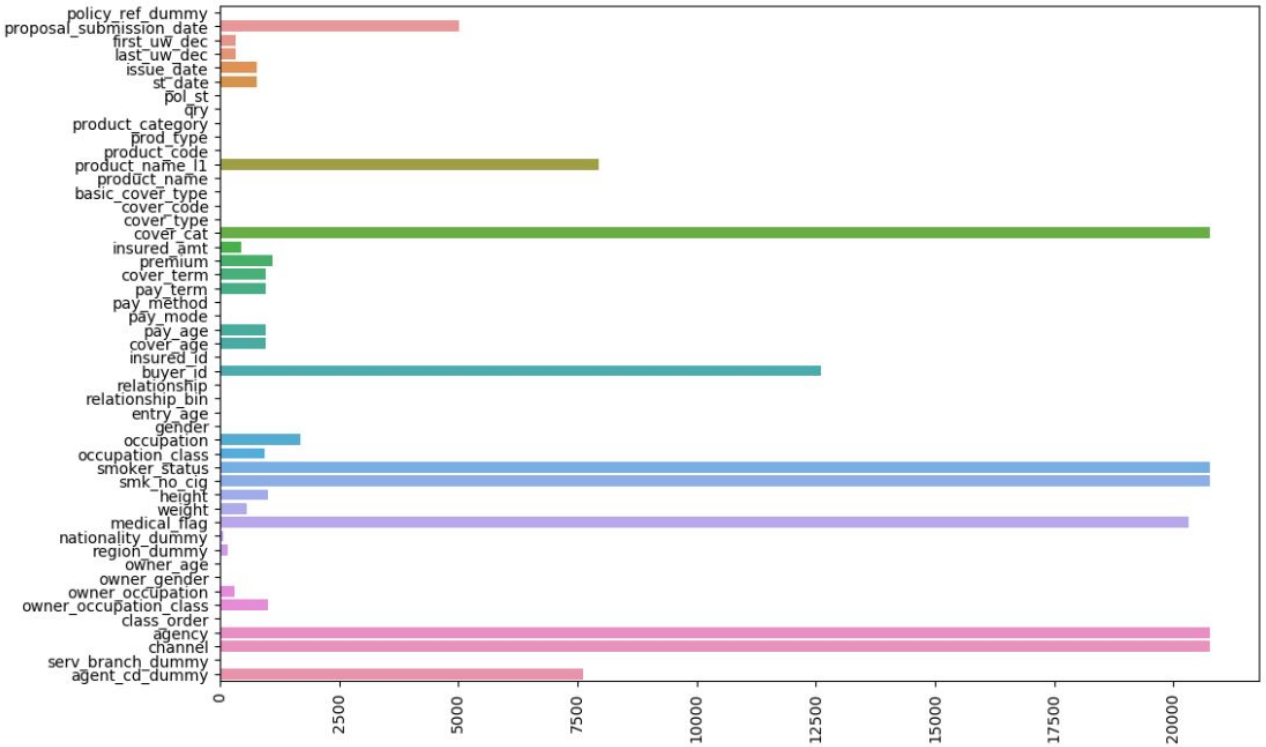
HEIGHT_CM	WEIGHT_KG
20752.000000	20752.000000
147.583317	50.987037
115.822525	22.846167
0.000000	0.000000
147.000000	40.000000
157.000000	54.000000
165.000000	65.000000
5608.000000	200.000000

INSURED_AMT	PREMIUM
2.075200e+04	2.075200e+04
5.030884e+05	1.905721e+05
1.059495e+06	6.337166e+05
0.000000e+00	-5.000000e+05
1.075278e+05	7.363900e+02
2.475144e+05	2.578392e+04
5.000000e+05	1.500000e+05
3.125000e+07	2.500000e+07

Missing Data

01

Several features are missing almost 80 - 90% of their information

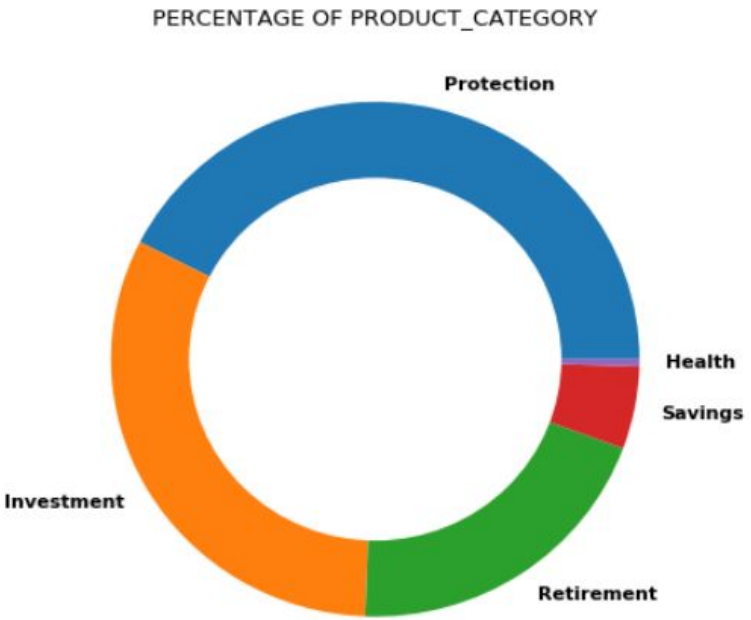


Buyers Behavior

02

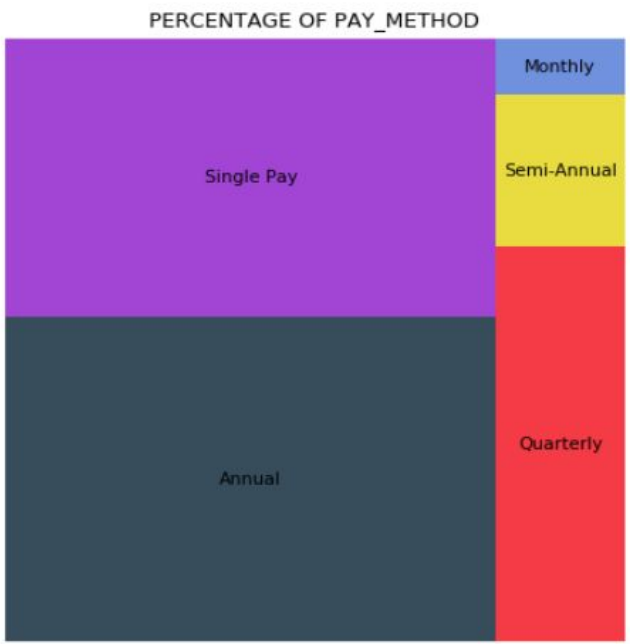
It’s clear to see that most people buy for Investment and Protection purposes

Client Implications:
Needs to upsell Retirement, Health and Savings



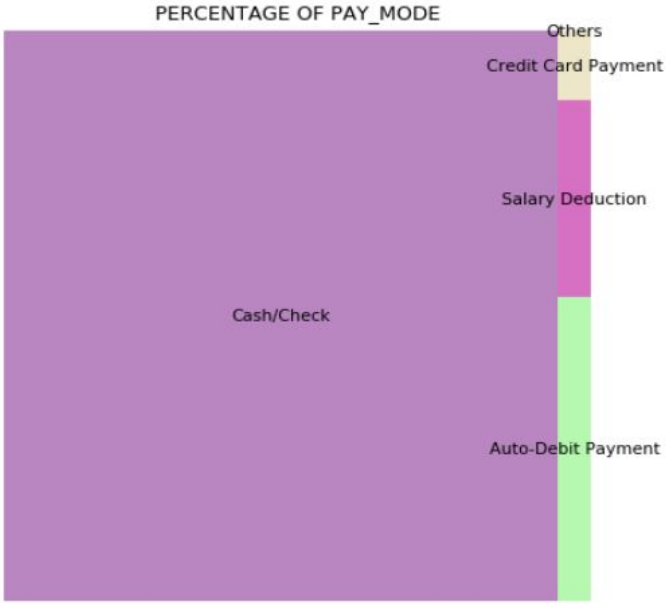
Buyers Behavior

02



Most people prefers to do Annual Pay (in this dataset) and Single Pay, and mostly in Cash/Check

Client Implications:
Needs to re-evaluate the process for other mode of payments and channel

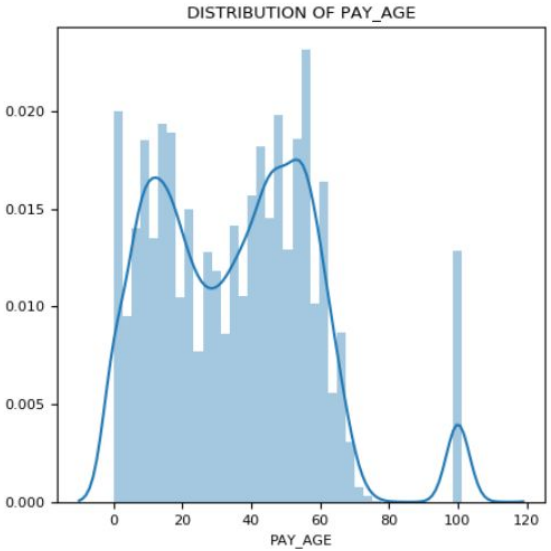


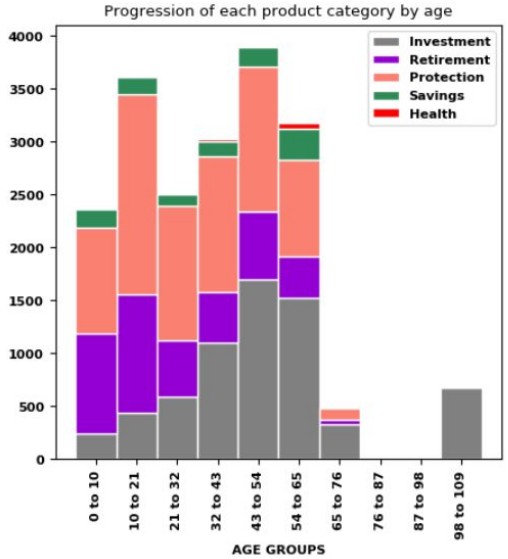
Trend analysis

We observes two peaks for young adults (High percentage of Protection) and middle age customers (Highest percentage of Investment)

Retirement decreases with age while Health is only a small portion

This is potentially due to: disposable income? Fear of the future?



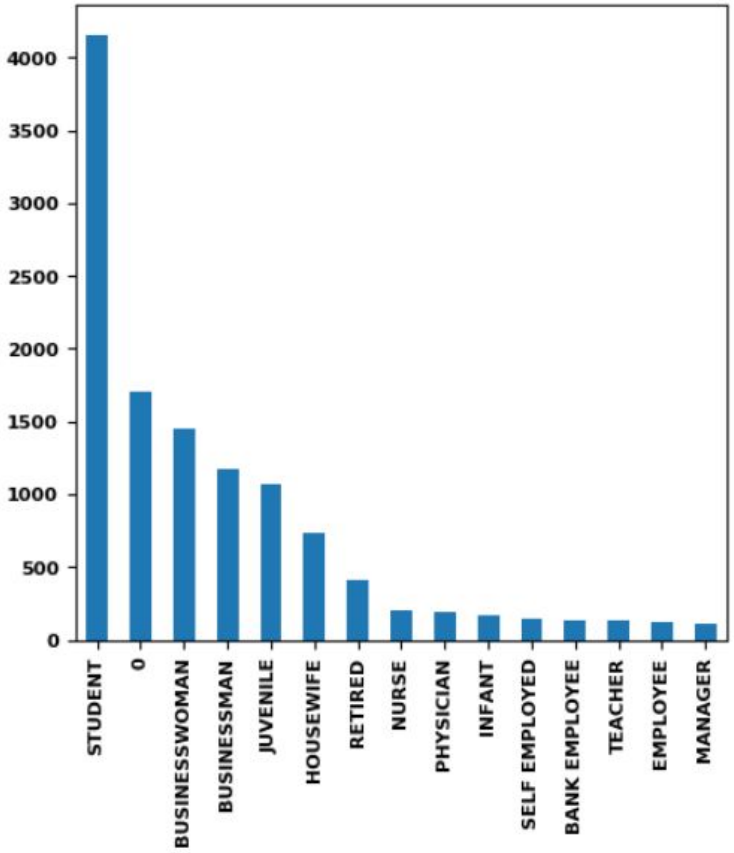


Trend analysis

This is the top 15 of the occupations recorded in the data (forsaking spelling mistakes)

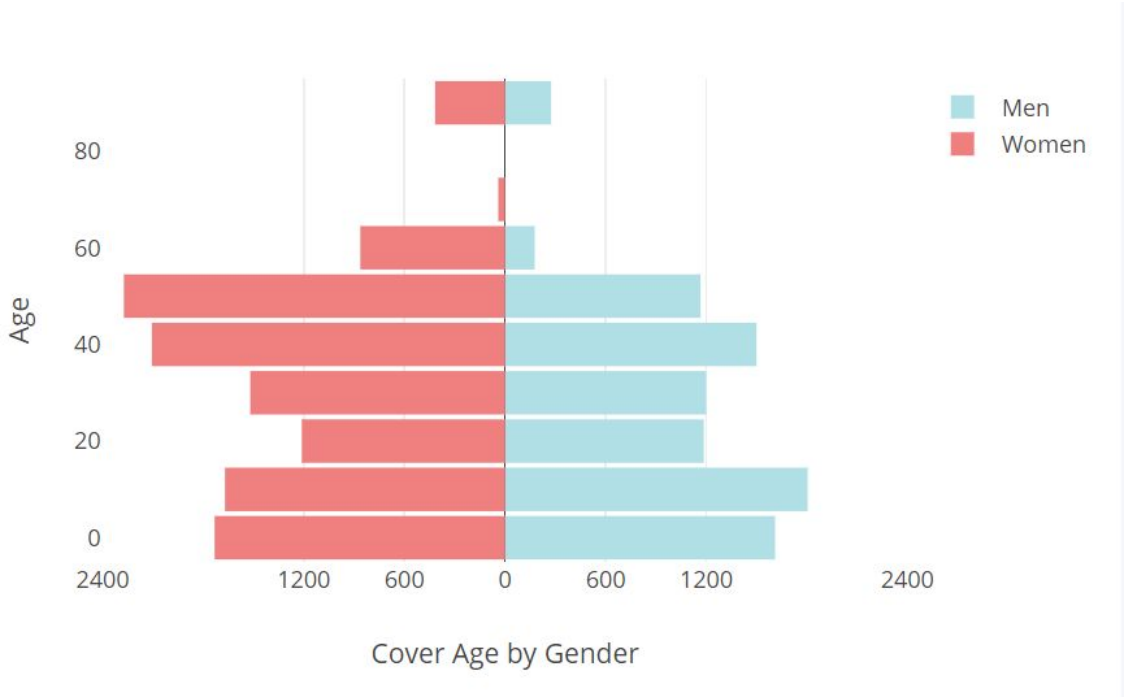
Students are a large group that are insured for while businessmen and women follow suit

Upselling measures need to target the other classes which have much lower count



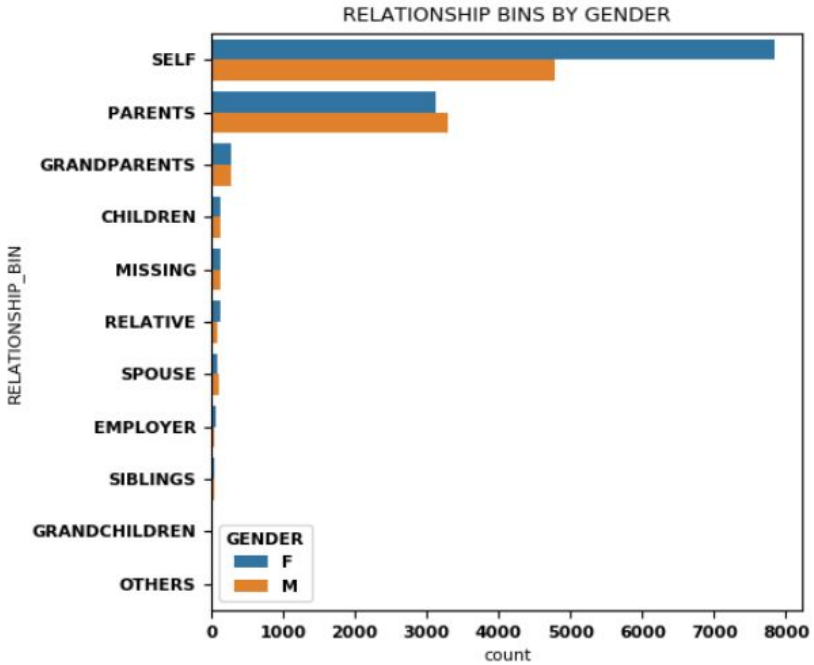
Gender Study

Although the two peaks remain the same, there is considerable difference when it comes to the number of buyers in each gender group



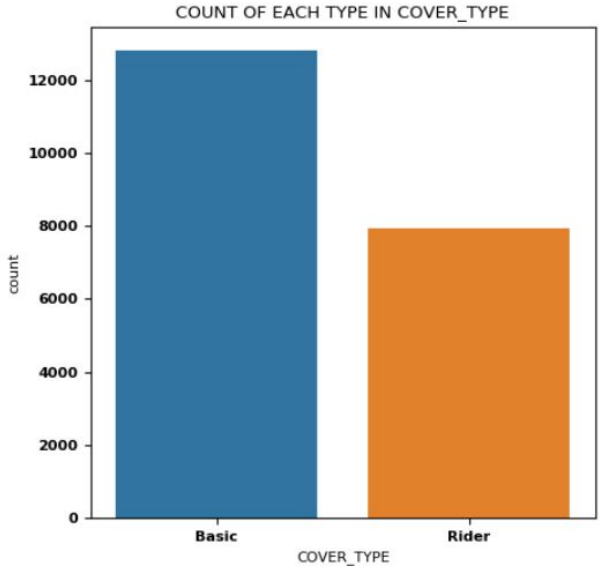
Gender Study

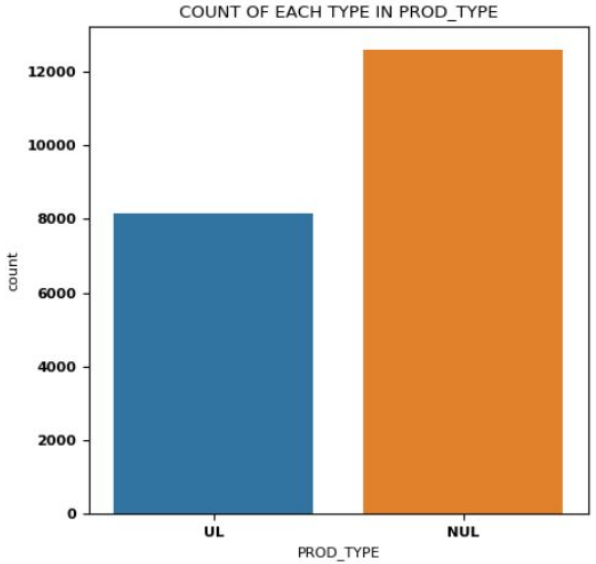
This also reflects in the difference between the two groups when it comes down to who to buy insurance for



Imbalanced classes

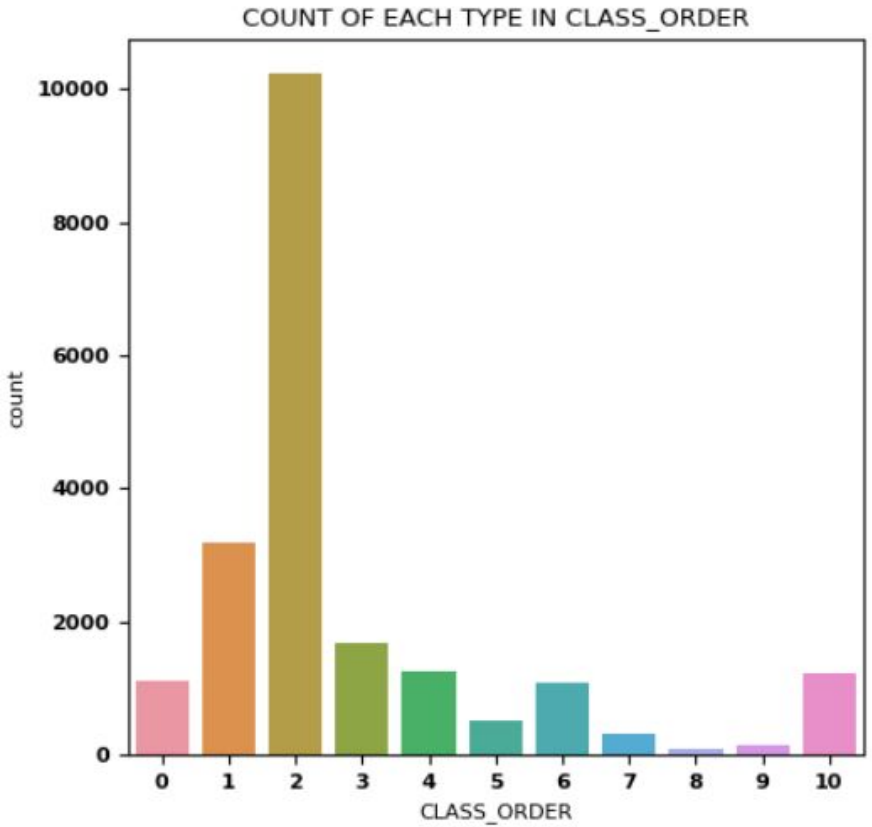
A few features suffer from this problems, which might affect the results





Imbalanced classes

A few features suffer from this problems, which might affect the results



Preliminary Observations

The buyers in this dataset do differ from each other in terms of age, buying habits and behavior. From the business point of view, the first step would be deep diving into what constitute these behaviors and try to address them.

As the data is not cleansed, the following steps were taken to ensure a quality feed to the machine learning model.

01 | Fill in missing values with 0 (Instead of imputing them)

02 | Convert height and weight into the same unit

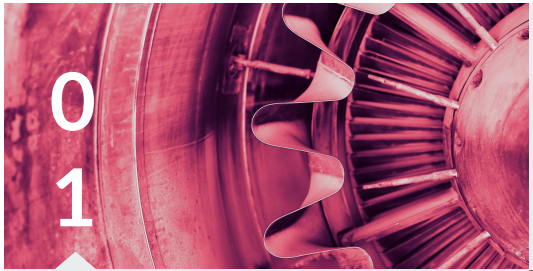
03 | Convert some columns to datetime type

04 | Standardize strings

05 | Create dummy for categorical variables, aware of dummy variable trap

06 | Feature engineering: ANP, RFM and Rating dataframe

Process



Explainable Model

Create a model that's explainable and applicable to non-technical audience

Descriptive Statistic

Drive insights through basic observation of the data. Create baseline model

Machine Learning Model

Using Machine Learning algorithms to overcome limitations of traditional models

Assumptions

I do not have the domain knowledge in this particular enterprise segment, therefore I have made certain assumptions that work with normal business models

01 | There are several repetitions in each of the occupation bins, this is intentional, accurate and will not affect the result

02 | For the RFM model used for clustering, recency is calculated from issue date

03 | For the Recommendation Engine, propensity to purchase a certain product category is treated as ‘rating’

04 | Missing medical flags are treated as no medical flags

05 | There is no real ordinal relationship in Class Order

06 | Purchase history is not influenced by Premium paid (basis for Apriori and Collaborative Filtering)

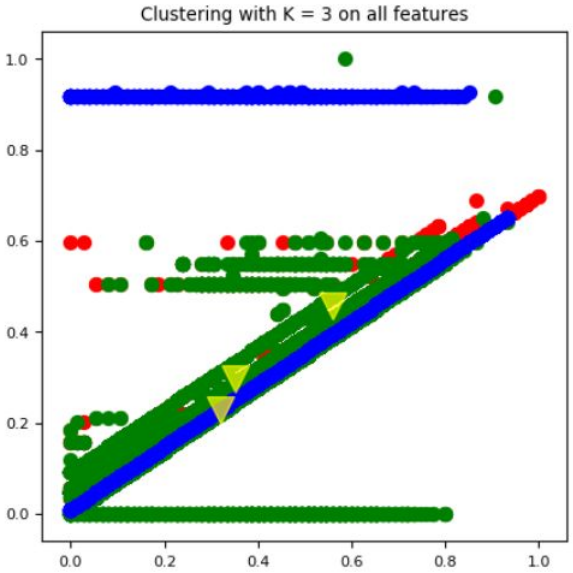
Proposed solution

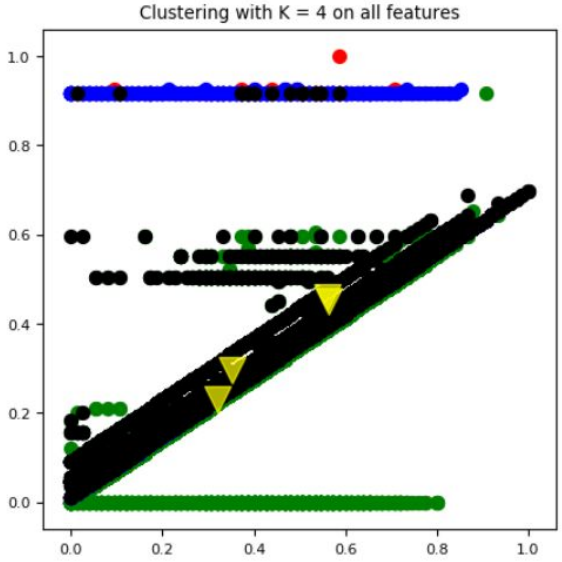
Find customer segmentation via clustering

Clustering
Whole Dataset

One of the challenges I face while exploring the customer segments is that running different clustering methods on the whole feature data give results that are not too clearcut. It is much harder to visualize and talking about what delineate them.

Additionally, I lack the domain knowledge to justify the model's complexity



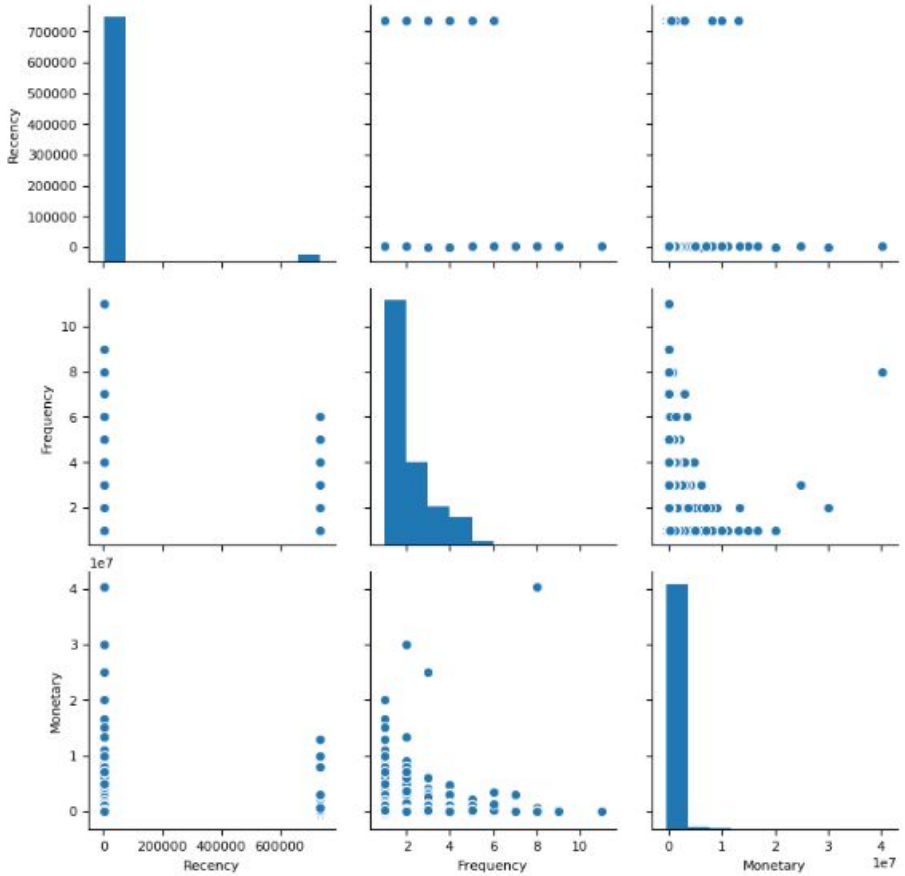


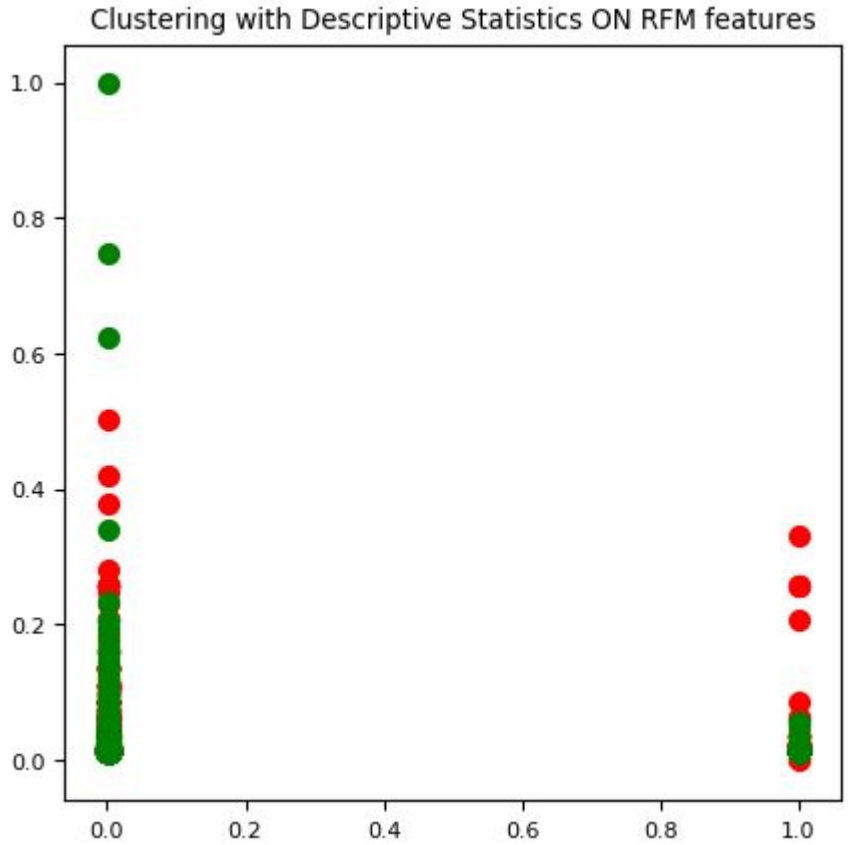
Clustering

RFM Model

Working on the RFM Model, it is much clearer and easier to communicate to the stakeholders why ML is a better choice than traditional model. Here it is clearer how there are clusters in the buyers population.

I chose the Recency-Frequency-Monetary model for establishing a baseline and creating a useful model





Baseline model

Based on percentile, the monetary vs recency features are chosen as an ad-hoc basis for visualization.

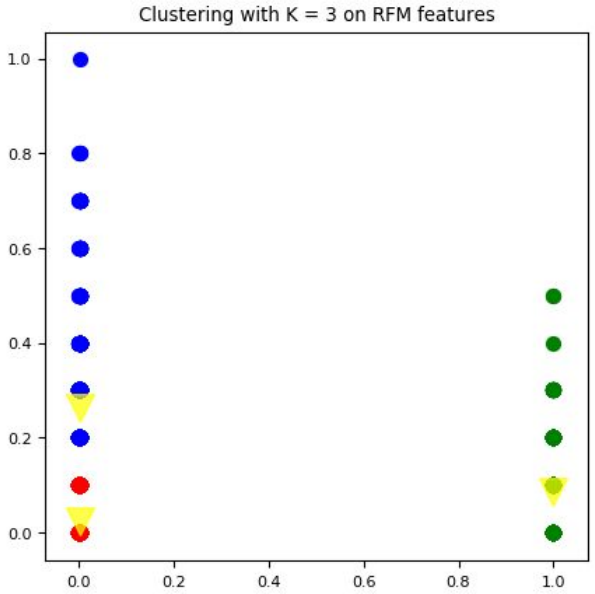
The different colors reprensent different clusters by percentile

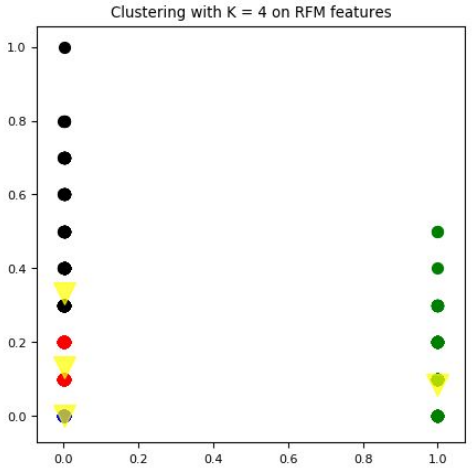
As evident from the graphs, there are no clear cluster established

K Means Clustering Model

Using K Means clustering algorithm on the RFM features, it is clearer that the clusters are established along with their centroids.

A limitations of KMeans is that due to its concept, a different result could be generated everytime. However, the clusters stay the same





K Means Clustering Model

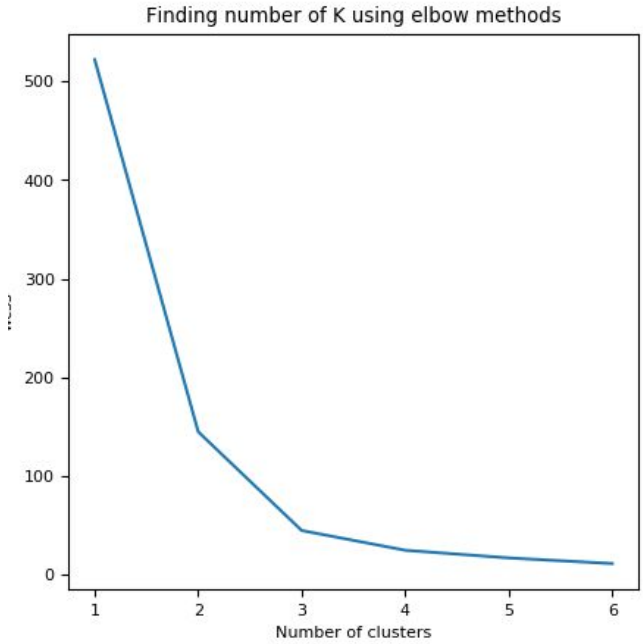
The criteria for each clusters are also shown

From this step, the team can deep dive on what are the major characteristics of each clusters based on regions, age, pay mode… etc and hence formulate their marketing efforts

```
==================================
Differentiating between clusters in method: kmeans_3
Group 0 Mean recency of group is 2047.3160372530958
Mean frecency of group is 1.249718554907379
Mean monetary of group is 360457.5193071339

Group 2 Mean recency of group is 2401.5511152416357
Mean frecency of group is 3.6319702602230484
Mean monetary of group is 148157.61887081803

Group 1 Mean recency of group is 737170.0
Mean frecency of group is 1.854219948849105
Mean monetary of group is 291272.0064961637


==================================
Differentiating between clusters in method: kmeans_4
Group 2 Mean recency of group is 1978.4450961669622
Mean frecency of group is 1.0
Mean monetary of group is 407900.90528168046

Group 0 Mean recency of group is 2263.1422958181306
Mean frecency of group is 2.315183833847881
Mean monetary of group is 196810.13940499572

Group 3 Mean recency of group is 2531.4927113702624
Mean frecency of group is 4.321671525753159
Mean monetary of group is 145101.60703595728

Group 1 Mean recency of group is 737170.0
Mean frecency of group is 1.854219948849105
Mean monetary of group is 291272.0064961637
```

Elbow method

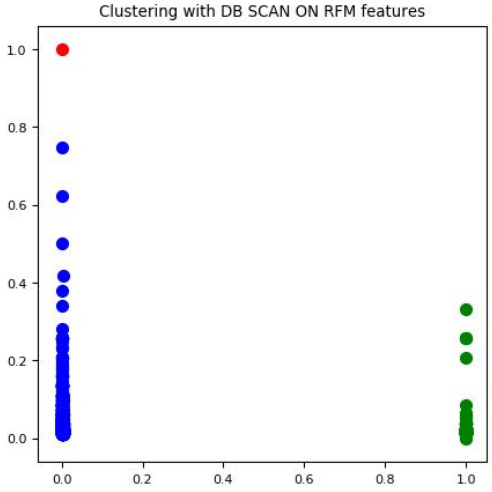
Minimizing the WCSS is one of the main methodology used to determine the number of centroids or K

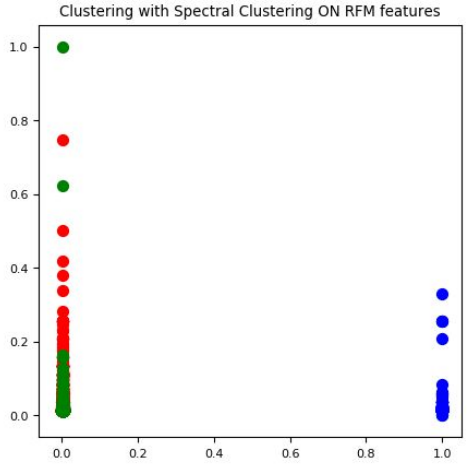


DB SCAN
Spectral Clustering

I have also tried out the other clustering methods. DB SCAN works better with denser clusters while Spectral clustering is a more novel method.

Again, how useful are the clusters depend on the business perspective and use case.





Results

The main metrics for evaluation are Silhouette Score for the clusters

KMEANS

0.80

K = 4 improves the score compared to K = 3

DB SCAN

0.89

DB Scan is the best performing model

SPECTRAL CLUSTERING

0.76

Spectral clustering underperforms

Proposed solution

Recommendation for Individual Customers

Aggregate
Purchase History

The category groups of each user historical purchases are combined into one single row of record.

We are using user-based collaborative filtering for this solution

INSURED_ID	PRODUCT_CODE	PRODUCT_CATEGORY
PART_ID_10	[2174]	Retirement
PART_ID_10000	[2163]	Investment
PART_ID_10001	[2163]	Investment
PART_ID_10002	[2413, 1543]	Retirement,Protection
PART_ID_10003	[2413, 1476, 1537]	Retirement,Protection,Protection
PART_ID_10004	[2304, 2313, 2332, 2319]	Retirement,Protection,Protection,Protection
PART_ID_10005	[2413, 1476, 1543]	Retirement,Protection,Protection
PART_ID_10006	[2413]	Retirement
PART_ID_10007	[1460]	Retirement
PART_ID_10008	[1465, 1481, 1541]	Savings,Protection,Protection

Apriori model

Using ideas such as lift, support, frequency of each category that are usually bought together, we could project a more macro decision for cross-selling

Such is not individualized solution but rather, the general trend of the buyers as a whole

	support	itemsets	length
5	0.000244	(Health, Investment)	2
6	0.005522	(Protection, Health)	2
7	0.002680	(Retirement, Health)	2
8	0.002030	(Health, Savings)	2
9	0.076011	(Protection, Investment)	2
10	0.007065	(Retirement, Investment)	2
11	0.003005	(Investment, Savings)	2
12	0.224622	(Retirement, Protection)	2
13	0.063343	(Protection, Savings)	2
14	0.041335	(Retirement, Savings)	2
15	0.000162	(Protection, Health, Investment)	3
16	0.000244	(Retirement, Health, Investment)	3
17	0.002599	(Retirement, Protection, Health)	3
18	0.002030	(Protection, Health, Savings)	3
19	0.000162	(Retirement, Health, Savings)	3
20	0.003654	(Retirement, Protection, Investment)	3
21	0.001787	(Protection, Investment, Savings)	3
22	0.002193	(Retirement, Investment, Savings)	3
23	0.023632	(Retirement, Protection, Savings)	3
24	0.000162	(Retirement, Protection, Health, Investment)	4
25	0.000162	(Retirement, Protection, Health, Savings)	4
26	0.000975	(Retirement, Protection, Investment, Savings)	4

	1	2	3	4	5
Health	Health	Savings	Protection	Retirement	Investment
Investment	Investment	Protection	Retirement	Savings	Health
Protection	Protection	Retirement	Savings	Health	Investment
Retirement	Retirement	Protection	Savings	Health	Investment
Savings	Savings	Protection	Retirement	Health	Investment

	Health	Investment	Protection	Retirement	Savings
Health	1	0.00281114	0.156664	0.041304	0.182044
Investment	0.00281114	1	0.104131	0.017485	0.0153525
Protection	0.156664	0.104131	1	0.540787	0.363784
Retirement	0.041304	0.017485	0.540787	1	0.241497
Savings	0.182044	0.0153525	0.363784	0.241497	1

Purchase Chain

Using Cosine Similarity, the ordinality of preference is established.

At this step, the team can already choose the next choice to cross sell (notwithstanding the quantity bought for each category)

User-based Collaborative Model

Using past purchase and the pattern of other customers, the next few categories are predicted on each single customers.

What has been bought before would not be recommended again.

	INSURED_ID	FIRST CHOICE	SECOND CHOICE	THIRD CHOICE	FOURTH CHOICE
0	PART_ID_10	Protection	Savings	Investment	Health
1	PART_ID_10000	Protection	Retirement	Savings	Health
2	PART_ID_10001	Protection	Retirement	Savings	Health
3	PART_ID_10002	Investment	Savings	Health	NaN
4	PART_ID_10003	Investment	Savings	Health	NaN

Other User-based Collaborative Model

Using MSE and RMSE as basis for evaluation, the model experimented with are KNN (for similar customers) and SVD

Again, only the quantity bought and habits of other users are taken into account.

```
Computing the cosine similarity matrix...
Done computing similarity matrix.
RMSE: 0.9078
Computing the cosine similarity matrix...
Done computing similarity matrix.
RMSE: 0.9255
Computing the cosine similarity matrix...
Done computing similarity matrix.
RMSE: 0.9121
```

```
Evaluating RMSE, MAE of algorithm SVD on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.9173  0.9102  0.9125  0.9158  0.9142  0.9140  0.0025
MAE (testset)     0.7838  0.7832  0.7797  0.7871  0.7819  0.7831  0.0024
Fit time          2.47    2.37    2.38    2.38    2.48    2.42    0.05
Test time         0.06    0.06    0.06    0.06    0.06    0.06    0.00
```


Deliverables

Output

The deliverable package will include the following:

01 | Presentation Deck

02 | Write-up

03 | Customer Segmentation: Each columns represent a cluster label generated from a method

04 | Cross-sale Recommendation: Apriori Information file

05 | Collaborative Filtering output: User-based

Proposal

- From the descriptive statistics point of view, it is possible to drive upsale for agegroups that are the bridge between the young adult and middle age. This is because they are interested in investment and savings, and have the disposable income to do so.
    - There is also a gap between gender for the product lines purchased and thus, we can also create customer segmentation based on age and try to drive more male customers.
    - From the apriori results, it is clear that many packages are frequently bought together. It thus could be marketed or bundled in such a way that's more likely to be received by the customers themselves (such as family bundle)

Proposal

- From the Data Science perspectives, customers could be segmented by their frequency and monetary value
    - As such, with the labelled clusters, the company can deep dive into the characteristics of each group that goes beyond the amount of premium they pay. A marketing solution could be tailored to each of these that go beyond the price tag but also how they cater to the specific needs of these groups
    - Many customers could be encouraged to try out less popular packages like saving, but perhaps with information on the effectiveness of the similar buyers in the same group.
    - Individualized recommendation results could be coupled with income and occupation information to make the packages more appealing to the customers

Limitations + Next Step

Data Manipulation

More feature engineering

Needs to acquire knowledge to create meaningful features while reducing the complexity and increasing accuracy

Demographic-based Collaborative Filtering

Utilize more features from dataset

Data is manipulated and cleaned however, only the rating and historical purchase features were used. Woul d be interesting to see more hybrid model for better customization

Segmentation

Increase Model Complexity

Go beyond RFM model
Utilize all other features to create more clusters

User-based Collaborative Filtering

More Experimentation

Create more model, GridSearchCV to improve MSE, compare individual predictions, A/B Testing.
Potential to do product-based collaborative filtering

Thank you.