Customer Segmentation

Combining RFM Analysis and Clustering to Predict Profitable Customers

Alexander Campbell

Receipt Bank
University of Cambridge
Alan Turing Institue

EARL
ENTERPRISE APPLICATIONS OF THE R LANGUAGE

Problem statement

Customer Segmentation

Sales and marketing resources are finite and expensive

- **Who are our most/least valuable customers?**
- **How can we acquire new customers that resemble our most valuable?**

Need a way of segmenting our customer base into groups based on their value to the business

Method

1) RFM + 2) Clustering + 3) Classification

- Combine customer value analysis with data mining techniques

- Emphasis on
  - Workflow, techniques, R packages
  - Easily interpretable results

Method

1) RFM

Describe current customers historical purchase behaviour using 3 feature:

- **Recency** = When did the customer make their last purchase?
- **Frequency** = How often does the customer make a purchase?
- **Monetary value** = How much money does the customer spend?

'Customers that purchase in shorter time intervals in greater volumes at higher prices are more like to respond positively to future engagement and product offers'

Method

1) RFM

Use *dplyr* to split customers into quintiles (5 groups) for each R, F and M:

```
rfm_data <- rfm_data %>%
mutate(R = ntile(desc(Recency), 5),
F = ntile(Frequency, 5),
M = ntile(Monetary, 5))
```

- Customers in top 20% of recency are given a score of 5, the next 20% a 4, and so on.
- Concatenate R, F, and M quintiles and rank from 555 to 111

Method

1) RFM

Id	Recency	Frequency	Monetary
Bob	100	3	10.89
Alex	2	100	90.26

↓

Id	Recency	Frequency	Monetary	R	F	M
Bob	100	3	10.89	1	4	5
Alex	2	100	90.26	3	3	5

↓

Id	Recency	Frequency	Monetary	RFM
Bob	100	3	10.89	145
Alex	2	100	90.26	335

Method

2) Clustering

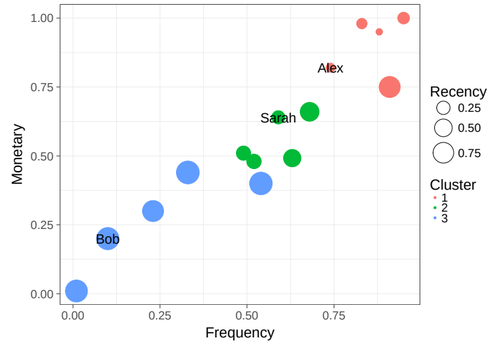
- Segment data into *k* clusters using algorithm from *cluster*

```
pam(rfm_data[, 2:4], k, metric = "euclidean", stand = True)
```

- For different values of *k* from 2 to 10 re-run the clustering algorithm > 500 times
- Find average Silhouette coefficient across each run
- Choose *k* with the highest average Silhouette coefficient

Method

2) Clustering



Plot results using *ggplot2* and label cluster centres

Method

3) Classification

- Add clusters to dataset

Id	RFM	Cluster
Bob	122	3
Alex	555	1
Sarah	335	2

- Introduce customer attributes

Id	RFM	Cluster	Country	Number employees
Bob	122	3	UK	10
Alex	555	1	Canada	60
Sarah	335	2	US	40

Method

3) Classification

- Build classification tree using *rpart*:

```
rfm_tree <- rpart(Cluster ., data = rfm_data)
```

- Plot tree to visualise classification rules:

```
rpart.plot(rfm_tree)
```

Results

Building a strategy

- Most valuable customers fall into cluster 1
  - Typified by Alex with R F M
  - More likely to be from either USA or Canada and have than 50 employees
- Strategy
  - Focus marketing and onboarding efforts on large US and Canadian customers
  - Keep recency and frequency of purchase as low as possible

Results

Building a strategy

- **Humanise the data by adding characters to the RFM quintiles**

RFM quintile	Character	1	2	3
R>3, F>3, M>3	Superstar	563	77	0
R<3, F<3, M>3	Churn Risk	10	100	340
R=3, F=3, M=3	Safe Bet	20	200	14

- Strategy
  - 'Churn Risks' start to appear in cluster 1
  - Discount price and engage customer ↓ recency and ↑ frequency

Results

Building a strategy

mixpanel Appboy

salesforce HubSpot

Conclusion

About 20% of your customers produce 80% of your sales

- **Customer Segmentation**
  - **RFM** $\Longrightarrow$ quantify value
  - **Clustering** $\Longrightarrow$ discover groups
  - **Classification** $\Longrightarrow$ differentiate & predict
- Four R packages
  - *dplyr*
  - *cluster*
  - *ggplot2*
  - *rpart*
- Build data driven strategies
- **Start small, go big, scale fast**

Questions

Questions?

ajrc4@cam.ac.uk

Projects?