Ford Ka Case

Wasif Mukadam

Executive Summary

Fords decision to develop a small car the Ka was in response to the environmental changes in the European market. Prior to the launch of their new model it is essential to correctly segment and target the market for Ford Ka to fully optimize their sales by targeting potential customers. The market research accumulated two vast types of data sets; a demographic standard approach and a psychographic; more personal approach to better understand their customers and competitors in the small car market. Predictive and exploratory analysis can be used on the datasets to provide key information on customer segmentation, targeting for Ford Ka and help determine the target market for the new product.

Definition of the Problem: In the 1980s and early 1990s there was a significant development in the car industry which affected the French car market. A series of environmental and demographic changes such as smaller parking areas in larger cities; lower fuel consumption as a result of added tax accounting in French regions made smaller cars look more appealing. The average household size reduced to less than three, more women began to work and many more factors that resulted in a smaller car being more attractive and feasible for everyday life. These drastic changes in peoples' everyday lives created a strong demand for smaller cars and brought about a new wave of buyers into the market. Ford needed to tackle this new desire of a smaller car fast and efficiently as the older traditional models were no more in "fashion". Ford used a market research team to understand the segments among the population and their competitors in the small car market.

Description of Data: Predictive and explanatory analysis was to be conducted on demographic and psychographic data to identify potential target customers. The two datasets were merged using a common column name "Respondent.Number". The as.factor function implemented in the lapply function was used to factor columns 3 to 10. The preference column was manipulated to form a binary column "pref.choosers"; where Ka choosers was given a 1 and Ka non-choosers and middle were given a 0. Customers who were undecided were given a 0 category in "pref.choosers" as they are less likely to purchase the Ford Ka car. As the data did not contain any missing values the data was split into a 75% training set and a 25% test set. Analysis of the data was then conducted after the data was handled.

Analysis and results: The first model conducted during the predictive analysis was conducted using a binary variable pref.choosers; the model indicated some statistically significant variables (figure 1). Gender2 (females) was seen as seen as statistically significant with a p-value of 0.00591; this indicates the odds of a female purchasing Ford ka is 2.79 times more. Gender2 being statistically significant aligns with the increase in women entering the working force in France during the 1990s; as well as women in general would prefer smaller cars as compared to larger ones. Q 20 and Q30 had the highest p-values among all the other questions. Q20 suggests people should look at their ability to make purchases based on their earnings and savings; Q20 had a negative coefficient which indicates the odds of these people choosing Ford Ka decreases by 0.63. Q30 on the other hand had a pe-value of 0.00411 and described owning a masculine car is important; this may be due to a larger population of people interested in cars are men. Therefore, the odds of people within Q30 choosing Ford ka increases by 1.69. Other variables such as FirstPurchase2, ChildCat1, and Q7 were seen as significant with a positive impact on being a Ford ka chooser. This may be due to people purchasing the car may already on a Ford brand and are comfortable with the products they release. Having only one child allows for a smaller car as the family size is small. Q7, people want a reliable car with no problems and longer lasting value.

Second, SVM was conducted; each coefficient's weight was used to determine its significance. Among all the factors, Q39 weighed the most at 0.1883, as smaller cars take up less space in traffic therefore, mare people would want a smaller car. With a weighted average of 0.155, Q52, relationship with my car, came in second; people who “love” their cars tend to be very cautious and gentle going to extreme measures the car remains in good conditions. MaritalStatus had the third highest ranking; this may be because people who are single tend to want smaller cars as compared to married people who have plans for a family in the future. In fourth, FirstPurchase; as people who already own a Ford would be more likely to buy the Ford Ka car. It is interesting to note that gender was among the top 8 weights in SVM; this is consistent with the arguments made in logistic regression that women are more likely to purchase smaller cars due to joining the workforce (figure 3).

Clustering was the only model implemented for the exploratory analysis of the case. Three distinct clusters were observable in the plot. The clusters formed were based on the closest centers from each point (figure 5). The three clusters indicated the most choosers were found in cluster 3 with pref.choosers centers at 49.5% while clusters 1 and 2 were 44.6% and 43.6% respectively. A boxplot of the densest cluster (cluster 1) indicated MaritalStatus, ChildCat, and income were most significant from the demographic data. This was in line with the observations from logistic and SVM. However, income was the only variable significant in clusters and not the previous two models; this could be because people with higher incomes are more likely to purchase a newer car over those with an average or low income. Q17, Q39, Q40 - 44 were most significant among the psychographic data (figure 6). Q17 may be people who want cars that are fast or fulfill the purpose of going from point A to B, Q39 aligns the discussion form SVM. Q40 to 44 is a variety of preferences from personally liking smaller cars to manufactures not caring about customers' needs and wants.

Among the three models used in predictive and exploratory analysis SVM may be the most optimum model to focus on in deciding what factors to account for when targeting customers. It had a train and test accuracy of 0.736 and 0.556 respectively, with a train and test error of 0.264 and 0.444 respectively (figure 4). Logistic regression had a lot higher accuracies and errors as seen in figure 2. Although clustering did a good job in segmenting the various demo and psychographics, a boxplot was required to identify the lowest scores as Ford ka choosers. Even though clusters help in visualizing the segmented customers SVM was easier to interpret and determine significant variables.

Recommendations/Conclusion: Overall, Ford should focus on Gender2, MaritalStatus, ChildCat, Q39, Q7 and Q52 as they were all consistent variables among the three analytical techniques used. Income is also a good variable to investigate when segmenting the customers. Ford should advance with the environmental changes and accommodate the peoples wants for smaller more efficient, environmentally friendly cars (based on significant variables). Ford should consider releasing different styles of the Ka brands to accommodate different tastes and preferences in the population.

Appendix

```
> ford.ka.f <- formula(pref.choosers ~ Gender + FirstPurchase + ChildCat + Q5 +Q7 + Q12 + Q20 + Q29 + Q30 + Q34 + Q39 + Q49 + Q52)
> fit.logit.final <- glm(ford.ka.f, data = train.dta, family = "binomial")
> summary(fit.logit.final)

Call:
glm(formula = ford.ka.f, family = "binomial", data = train.dta)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8122  -0.9839  -0.3747   1.0259   2.2842

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.30949    2.28102   1.451  0.14681
Gender2         1.02716    0.37315   2.753  0.00591 **
FirstPurchase2  1.14437    0.50844   2.251  0.02440 *
ChildCat1       1.04834    0.46416   2.259  0.02391 *
ChildCat2       0.05971    0.43760   0.136  0.89146
Q5             -0.26576    0.14745  -1.802  0.07149 .
Q7              0.49633    0.19642   2.527  0.01151 *
Q12            -0.36596    0.15663  -2.336  0.01947 *
Q20            -0.45730    0.16733  -2.733  0.00628 **
Q29            -0.29035    0.16847  -1.723  0.08480 .
Q30             0.52576    0.18323   2.869  0.00411 **
Q34            -0.35030    0.17532  -1.998  0.04572 *
Q39            -0.29483    0.15904  -1.854  0.06377 .
Q49            -0.35373    0.16697  -2.118  0.03413 *
Q52             0.21410    0.12518   1.710  0.08720 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 258.59  on 186  degrees of freedom
Residual deviance: 213.80  on 172  degrees of freedom
AIC: 243.8

Number of Fisher Scoring iterations: 4
```

Figure 1: glm of stepwise feature selection

```
> logit.pred.prob.train <- predict(fit.logit.final, newdata = train.dta, type = "response")
> logit.pred.train <- ifelse(logit.pred.prob.train > 0.5, 1, 0)
> mean(logit.pred.train != train.dta$pref.choosers)
[1] 0.2834225
> confusionMatrix(as.factor(logit.pred.train), as.factor(train.dta$pref.choosers), mode = "everything")
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 74 28
         1 25 60

               Accuracy : 0.7166
                 95% CI : (0.6462, 0.7799)
    No Information Rate : 0.5294
    P-Value [Acc > NIR] : 1.291e-07

                  Kappa : 0.4301

 Mcnemar's Test P-Value : 0.7835

            Sensitivity : 0.7475
            Specificity : 0.6818
         Pos Pred Value : 0.7255
         Neg Pred Value : 0.7059
              Precision : 0.7255
                 Recall : 0.7475
                     F1 : 0.7363
             Prevalence : 0.5294
         Detection Rate : 0.3957
   Detection Prevalence : 0.5455
      Balanced Accuracy : 0.7146

       'Positive' Class : 0
```

```
> logit.pred.prob <- predict(fit.logit.final, newdata = test.dta, type = "response")
> logit.pred <- ifelse(logit.pred.prob > 0.5, 1, 0)
> mean(logit.pred != test.dta$pref.choosers)
[1] 0.4126984
> confusionMatrix(as.factor(logit.pred), as.factor(test.dta$pref.choosers), mode = "everything")
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 20 11
         1 15 17

               Accuracy : 0.5873
                 95% CI : (0.4562, 0.7099)
    No Information Rate : 0.5556
    P-Value [Acc > NIR] : 0.3535

                  Kappa : 0.1761

 Mcnemar's Test P-Value : 0.5563

            Sensitivity : 0.5714
            Specificity : 0.6071
         Pos Pred Value : 0.6452
         Neg Pred Value : 0.5312
              Precision : 0.6452
                 Recall : 0.5714
                     F1 : 0.6061
             Prevalence : 0.5556
         Detection Rate : 0.3175
   Detection Prevalence : 0.4921
      Balanced Accuracy : 0.5893

       'Positive' Class : 0
```

Figure 2: Training set (Left) and test set (right) confusion matrices

```
> print(w)
        Q39         Q52 MaritalStatus FirstPurchase         Q49          Q7         Q16      Gender
0.188308317 0.155467697   0.151209121   0.148185807 0.134398558 0.122284717 0.118889117 0.118077500
        Q12   IncomeCat           Q59           Q50         Q55         Q34    ChildCat         Q30
0.116209837 0.113015374   0.106469206   0.105191067 0.100439753 0.099339157 0.092202953 0.091994063
        Q24         Q47           Q19           Q25          Q3         Q44         Q33         Q42
0.091694056 0.090290105   0.086623928   0.078105299 0.073205772 0.073039400 0.072658209 0.071059031
        Q58          Q4           Q48           Q37          Q6         Q21         Q10         Q18
0.070547541 0.069984520   0.069210439   0.068116490 0.059963420 0.059220069 0.058430097 0.057282815
         Q8         Q60           Q26           Q32         Q29      AgeCat         Q35          Q9
0.054568081 0.052736528   0.052483357   0.052221080 0.051600262 0.050964563 0.049221338 0.046075416
        Q13         Q62           Q53           Q43         Q22         Q36         Q28          Q2
0.039110800 0.038856644   0.036813693   0.034900301 0.034854792 0.032687036 0.030818753 0.029291748
        Q11         Q15           Q17           Q38         Q51         Q31         Q61         Q20
0.027345757 0.024463581   0.023816266   0.023055189 0.022789427 0.022051095 0.021558512 0.020970540
        Q14         Q57           Q40           Q56         Q46         Q27         Q23         Q54
0.015896228 0.014779249   0.013941378   0.013444888 0.012005216 0.011398894 0.010708794 0.009759786
        Q45          Q1            Q5           Q41
0.009005667 0.006338627   0.006240195   0.005881073
```

Figure 3: SVM weights for each of the variables

```
> confusionMatrix(as.factor(Ford.svm.train.predict),as.factor(trainset$pref.choosers))
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 61 19
         1 28 70

               Accuracy : 0.736
                 95% CI : (0.6648, 0.7991)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 1.144e-10

                  Kappa : 0.4719

 Mcnemar's Test P-Value : 0.2432

            Sensitivity : 0.6854
            Specificity : 0.7865
         Pos Pred Value : 0.7625
         Neg Pred Value : 0.7143
             Prevalence : 0.5000
         Detection Rate : 0.3427
   Detection Prevalence : 0.4494
      Balanced Accuracy : 0.7360

       'Positive' Class : 0

> #tree test error prediction
> Ford.svm.test <- predict(Ford.svm,testset)
> Ford.svm.test.predict<-ifelse(Ford.svm.test>0.5,1,0)
> Ford.svm.test.error <- mean(Ford.svm.test.predict!=testset$pref.choosers)
> Ford.svm.test.error
[1] 0.4444444
> confusionMatrix(as.factor(Ford.svm.test.predict),as.factor(testset$pref.choosers))
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 23 10
         1 22 17

               Accuracy : 0.5556
                 95% CI : (0.4336, 0.6728)
    No Information Rate : 0.625
    P-Value [Acc > NIR] : 0.90864

                  Kappa : 0.1293

 Mcnemar's Test P-Value : 0.05183

            Sensitivity : 0.5111
            Specificity : 0.6296
         Pos Pred Value : 0.6970
         Neg Pred Value : 0.4359
             Prevalence : 0.6250
         Detection Rate : 0.3194
   Detection Prevalence : 0.4583
      Balanced Accuracy : 0.5704

       'Positive' Class : 0
```

Figure 4: SVM confusion matrices for normalized train and test set

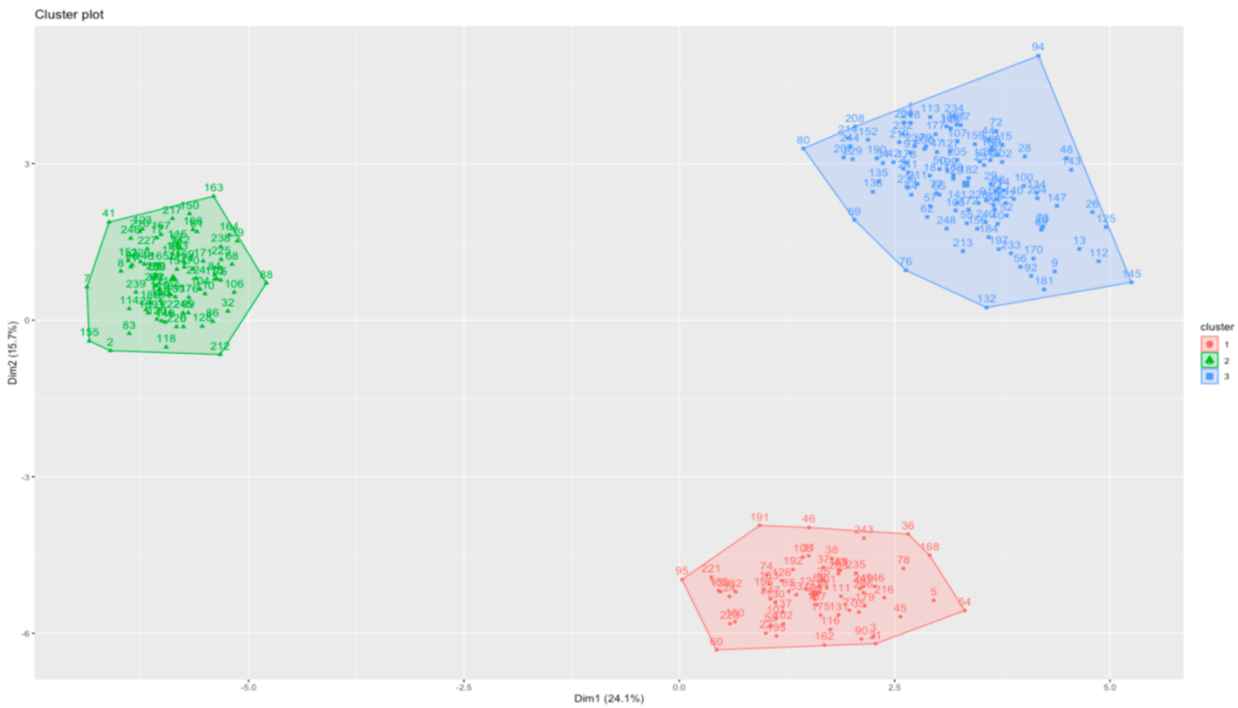

Figure 5: Clustering of normalized data

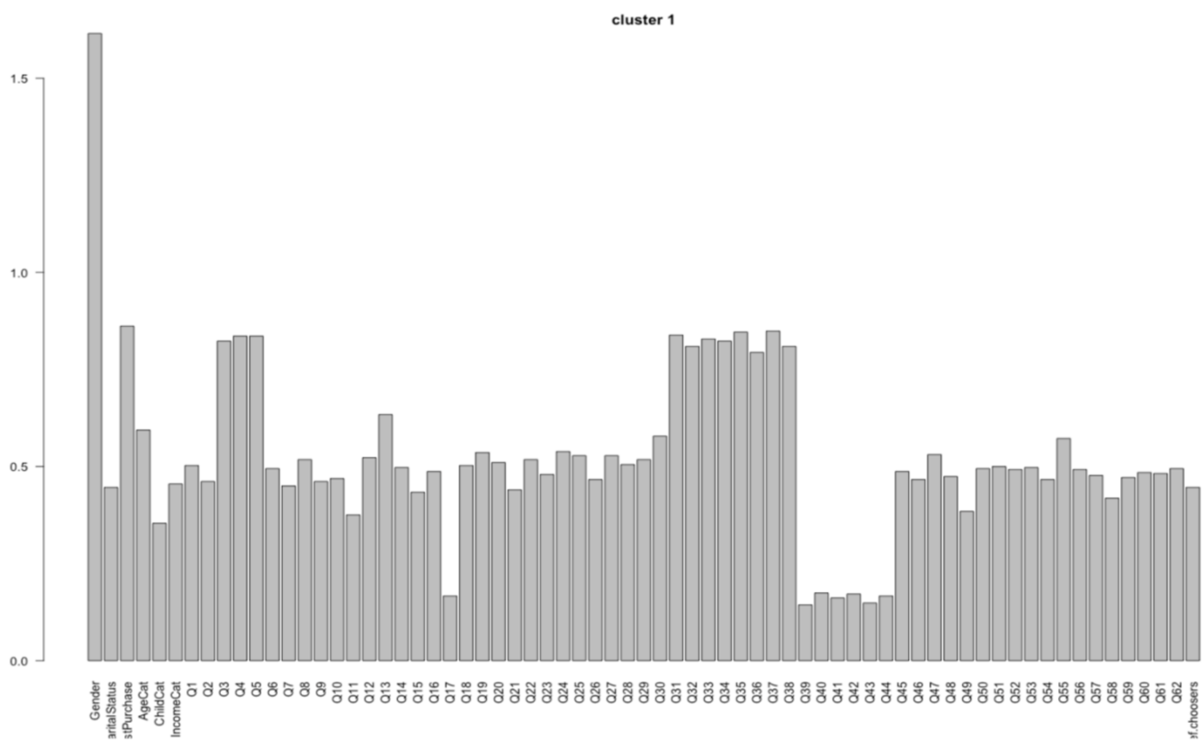

Figure 6: Bar chart representing the clusters