TEAM - 4

INDEX

1. Problem Statement
2. Objective
3. Dataset Available
4. Data Transformation
5. Feature Selection
    i. Boruta Feature Selection Algorithm
    ii. Forward Selection and Backward Elimination Algorithm
    iii. Random Forest Feature Importance Method
6. Model Building
7. Results
8. Conclusion
9. Annexure 1
10. Annexure 2

Problem Statement

The problem is regarding the analysis of survey data, for 10 popular deodorants, of a leading consumer packaged goods company. The survey questions are divided into two parts:

- Screener Questionnaire – These questions are to decide whether the person filling up the survey is relevant or eligible to complete it properly.
- Main Questionnaire – These questions are specific to a particular deodorant and are designed for its evaluation.

Objective

To develop a model that will predict the Instant Liking, along with its probability on a scale of 1-7, for the following deodorants: Deodorant B, Deodorant F, Deodorant G, Deodorant H and Deodorant J. Through this analysis, the company should be able to understand the factors/features that affect the likeability of each of the above products.

Dataset Available

The dataset consisted of 15217 rows combined for all 10 deodorants with 174 total variables including target variable, Instant Liking. For the required deodorants, individual no. of rows are:

- Deodorant B – 1521 rows, 75.08% Like
- Deodorant F – 1520 rows, 74.41% Like
- Deodorant G – 1520 rows, 75.06% Like
- Deodorant H – 1522 rows, 73.98% Like
- Deodorant J – 1522 rows, 74.44% Like

Data Transformation

There were a few discrepancies in the dataset:

- In both Q2 and Q7, the main questionnaire in the problem statement mentioned that symbol 1=negative and 2=positive, whereas in the data dictionary as well as in the data 1=positive and 2=negative. We **assumed data dictionary to be correct** and chose 1 for positive and 2 for negative respectively.
- In Q6 and Q7, only 24 list items were mentioned in main questionnaire, whereas 32 and 31 variables were given for these questions in the dataset respectively. Even the data dictionary mentioned just 24 and not 32 or 31. We **assumed data dictionary to be correct** and removed extra variables accordingly.
- Q8 was supposed to be a single choice question, but multiple choice values were there for many survey respondents. We **let this questions, as it is**.

We transformed the variables in the following ways:

- In Q2 of main questionnaire, 10 variables were present with 5 corresponding to 5 words and other 5 being their sentiment, as marked by the survey respondent itself. In order to capture the essence of Q2, we created a new variable which gave us an overall positive sentiment of these 5 words in a range of 0 to 5, where **0 means no** positive word said and **5 means all** positive words said.
- Q6 and Q7 of main questionnaire were complimentary to each other. In Q6 a survey respondent chose a few words, from an existing list, and in Q7 she marked sentiment of the words chosen by her in **Q6**. We removed 24 variables related to Q6 from the model, as all **information** in these 24 variables was **contained in** 24 variables of **Q7**.
- In Q7 of main questionnaire, **our assumption** was that not all characteristics can be present in the scent of a deodorant and by looking at the options mostly chosen in a particular deodorant by the survey respondents, one can **easily determine** the most important characteristics for that deodorant. **To further back our assumption**, among the 24 variables, we chose the top 10 most important features using Random Forest's feature importance function. And combined these **10 features** to give us an **overall sentiment** about characteristics of the scent of a deodorant.
- We used similar assumption in Q8, to reduce to 10 important variables, as in Q7.
  Multiple Correspondence Analysis - We have used **MCA** (Multiple Correspondence Analysis) to back our **assumption to variable selection** in Q7 and Q8. Using MCA we create **perceptual maps** on the features in each question before and after variable selection. The **relative distances** between points in the perceptual maps are proportional to the correspondence and association in the variables, whereas **overlapping** or very close points convey that the variables are closely associated. **Before feature selection** a large number of variables overlap however **after feature selection** the variables are distinct showing removal of closely associated variables thereby **reducing redundant information**. *(Related plots can be found in the Annexure 1)*
- In Q13 of screener questionnaire, 50 variables were given corresponding to each option in the question. We combined these into 1 variable, which gave us the total of all options chosen in Q13 i.e. count of the no. of distinct deodorants she has regularly worn.
- Q13A of screener questionnaire tells us about the deodorant she most often wear i.e. **her favourite deodorant**. So, if someone gives survey about a deodorant which is her favourite, as told by her in Q13A, that **increases the probability** of Instant Like it. If she doesn't mark the same deodorant, she has surveyed for, as her favourite, then this information becomes redundant. So, we replaced these entries in all the rows, where favourite is same as the survey deodorant by 1 and all other by 0.

Generally, in a survey, a respondent **cannot accurately differentiate** between being "Much too Strong" and "Somewhat too Strong" or between "Somewhat too weak" and "Much too weak". And thus, these ratings can almost be taken as equal. So taking this **assumption** into consideration, we re-scaled Q3, Q5, Q9, Q10 and Q13 in the following manner:

- In Q3, Q5, Q9 and Q10 of main questionnaire, rating 4 & 5 was replaced by 1 and rest by 0.
- In Q13 of main questionnaire, rating 5, 6 & 7 was replaced by 1 and rest by 0.

Feature Selection

We used the following feature selection techniques:

1. Boruta Feature Selection Algorithm
2. Forward Selection & Backward Elimination Algorithm
3. Random Forest Feature Importance Method

Boruta Feature Selection Algorithm

It is a feature selection algorithm that works as a wrapper algorithm around Random Forest. The **main goal** of the problem statement was **to give insights** about the features/factors that mostly affect the likeability of each deodorant. Hence, keeping this mind, we used Boruta Algorithm which **helps** us in **understanding** the mechanism related to our target variable i.e. Instant Liking, rather than just building a black box predictive model with good prediction accuracy. It follows an **all-relevant feature selection** method, where it captures all **features** which are in some circumstances **relevant** to the target variable. We used the Boruta package in R for its implementation.

Forward Selection & Backward Elimination Algorithm

Both of these algorithms are parts of wrapper method. Wrapper methods evaluate subsets of variables which allows, unlike filter approaches, to detect the possible interactions between variables. We used AIC value as benchmark for both forward and backward algorithms.

Akaike Information Criterion (AIC) - It is a **measure** of the **relative quality** of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Hence, AIC provides a **means for model selection**. It deals with the trade-off between the goodness of fit of the model and the complexity of the model.

Forward Selection - In this approach, one **adds variables** to the model **one at a time**. At each step, each variable that is not already in the model is tested for inclusion in the model. The most significant of these variables is added to the model, so long as it is decrementing the AIC-value, after getting introduced. Thus we begin with a model including the **variable** that is **most significant** in the initial analysis, and continue adding variables until none of remaining variables are "significant", when added to the model.

Forward selection has drawbacks, including the fact that each addition of a new variable may render one or more of the already included variables non-significant. **For this reason**, we also applied Backward Elimination.

Backward Elimination - Under this approach, one starts with fitting a model with all the variables of interest .Then the **least significant variable** is dropped on, which again get **decided by** the **decrement of AIC value** below earlier cut off, so long as it is not significant at our chosen critical level. We continue by successively re-fitting reduced models and applying the same rule until **all remaining variables** are **statistically significant i.e. no decrement possible in AIC value further**. Sometimes there maybe variables, which are dropped, but can be significant when added to the final reduced models.

The two main disadvantages of these wrapper methods are:

- The increase in overfitting risk when the number of observations is insufficient.
- The significant computation time when the number of variables is large.

Random Forest Feature Importance Method

Random Forest **works really well with categorical variables** and we used them to **rank** the predictor variables in **order of importance** in predicting the outcome. Individual decision trees intrinsically perform feature selection by selecting appropriate split points. This information can be used to measure the importance of each feature, the basic idea is, the **more often a feature is used** in the split points of a tree the more important that feature is. Also, **if the variable is not important**, then rearranging the values of that variable will not degrade prediction accuracy. Thus, the **relative rank** (i.e. depth) of a feature used as a decision node in a tree can be used to **assess the relative importance** of that feature with respect to predictability of the target variable.

Next we picked up the **top 25 most important** features from **55 consolidated features** from **RF feature importance** method, and used them to develop **Logistic Regression models** in **R**. Henceforth, the dataset was split for **training and validation**. The features selected were used to develop a model over the training set and the results were cross-validated on the validation set using **Logarithmic Loss**, as our **metric of evaluation.**

Model Building

The first and foremost part of building a model is selection of features. There are multiple methods already available for selection of features. There are 4 different logistic regression models available in R **Generalized Linear Models** that we used, **Simple GLM, Caret GLM, Bayesian GLM and Boosted GLM.** These 4 models were applied to the subsets of features, along with **grid search**, which we extracted from the feature selection processes as mentioned above. At the end of our methodology, we had **16 different models** (4 for each feature selection x 4 different logistic regression models) for each of the 5 deodorants and we used our metric **Logarithmic Loss function** to select the best model for each deodorant and then results are shown in the next section.

Logarithmic Loss (Model Evaluation Metric) –

This is an error metric which is used in measuring the performance of a **classification algorithm** which uses the **logarithmic cost function**. Log loss is used as a metric where we have to perform model selection and all the **models give a probability** ranging from definitely true(1) to equally true(0.5) and false(0). Log loss **measures** the **uncertainty of the probabilities** of your model by comparing them to the true labels. Log Loss **heavily penalises** classifiers that are confident about an **incorrect classification**.

Log-Loss function is used with $y_i \in \{0,1\}$ type of data with estimates $\hat{y}_i \in \{0,1\}$. Log-Loss function is defined as $logLoss = -\frac{1}{N}\sum_{n=1}^{N}[y_i log(p_i) + (1-y_i)log(1-p_i)]$ where p_i is the probability predicted by the model. This metric this has a range from 0 to infinity. Thus lower the value of log loss less is the uncertainty in the model.

Results (*lowest in 3 models, best model) (*Coefficient's of best model for each deodorant are shown in Annexure 2*)

Deodorant B –

1. **Top 10 features, Q7 -** q7_1 airy, q7_2 ambery,q7_5 cool, q7_11 floral, q7_13 fruity, q7_14 herbal, q7_18 powdery, q7_21 spicy, q7_23 pleasantly sweet, q7_28 woody
2. **Top 10 features, Q8 -** q8.1, q8.2, q8.5, q8.6, q8.8, q8.11, q8.12, q8.13, q8.19, q8.20
3. **Final features and log-loss table for Boruta Feature Selection Algorithm -** q2_all.words, q3_1.strength.of.the.Deodorant,q5_1.Deodorant.is.addictive,q7,q9.how.likely.would.you.be.to.purchase.this.Deodorant,q10.prefer.this.Deodorant.or.your.usual.Deodorant, Q13_Liking.after.30.minutes,q14.Deodorant.overall.on.a.scale.from.1.to.10,q4_2.attractive,q4_8.easy.to.wear,q4_11.for.someone.like.me,q4_13.high.quality,q4_16.memorable, q4_22.sophisticated, q4_23.upscale, q4_24.well.rounded, s13.2, s13a.b.most.often

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	**0.549***
Caret GLM	0.751
Bayesian GLM	1.2

4. **Final features and log-loss table for Forward Selection Algorithm -** q2_all.words, q4_12.heavy, q4_23.upscale, q4_13.high.quality, q5_1.Deodorant.is.addictive, q11.time.of.day.would.this.Deodorant.be.appropriate,s13a.b.most.often,q9.how.likely.would.you.be.to.purchase.this.Deodorant,q14.Deodorant.overall.on.a.scale.from.1.to.10, q7, Q13_Liking.after.30.minutes

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.572*
Caret GLM	0.733
Bayesian GLM	1.195

5. **Final features and log-loss table for Backward Elimination Algorithm –** q2_all.words, q4_1.artificial.chemical, q4_2.attractive, q4_3.bold, q4_5.casual, q4_6.cheap, q4_8.easy.to.wear, q4_9.elegant, q4_12.heavy, q4_13.high.quality, q4_15.masculine, q4_18.old.fashioned, q4_19.ordinary, q4_20.overpowering, q4_21.sharp, q4_23.upscale,q5_1.Deodorant.is.addictive, q7,q9.how.likely.would.you.be.to.purchase.this.Deodorant, q10.prefer.this.Deodorant.or.your.usual.Deodorant, q11.time.of.day.would.this.Deodorant.be.appropriate, Q13_Liking.after.30.minutes, q14.Deodorant.overall.on.a.scale.from.1.to.10, ValSegb, s10.income, s13a.b.most.often, s13b.bottles.of.Deodorant.do.you.currently.own

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.58*
Caret GLM	0.734
Bayesian GLM	1.186

6. **Final features and log-loss table for Random Forest Feature Importance Algorithm -** q4_24.well.rounded, q4_3.bold, q4_8.easy.to.wear, q4_2.attractive, q4_6.cheap,

q4_23.upscale, q4_22.sophisticated, q4_11.for.someone.like.me, q4_13.high.quality, q4_12.heavy, q4_21.sharp, q4_19.ordinary, q4_20.overpowering, q4_5.casual, q4_18.old.fashioned, q4_1.artificial.chemical, q5_1.Deodorant.is.addictive, q10.prefer.this.Deodorant.or.your.usual.Deodorant, q2_all.words, s13b.bottles.of.Deodorant.do.you.currently.own, ValSegb,q7,Q13_Liking.after.30.minutes,s10.income, q14.Deodorant.overall.on.a.scale.from.1.to.10

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.578*
Caret GLM	0.738
Bayesian GLM	1.165

Deodorant F –

1. **Top 10 features, Q7 -** q7_1 airy, q7_2 ambery, q7_5 cool, q7_6 Creamy, q7_11 floral, q7_13 fruity, q7_14 herbal, q7_18 powdery, q7_21 spicy, q7_28 woody
2. **Top 10 features, Q8 -** q8.2, q8.3, q8.4, q8.8, q8.10, q8.12, q8.14, q8.15, q8.16, q8.17
3. **Final features and log-loss table for Boruta Feature Selection Algorithm -** q2_all.words, q3_1.strength.of.the.Deodorant, q5_1.Deodorant.is.addictive, q7, q9.how.likely.would.you.be.to.purchase.this.Deodorant, q10.prefer.this.Deodorant.or.your.usual.Deodorant,Q13_Liking.after.30.minutes,q14.Deodorant.overall.on.a.scale.from.1.to.10, q4_2.attractive, q4_5.casual, q4_7.clean, q4_9.elegant, q4_11.for.someone.like.me, q4_17.natural, q4_22.sophisticated, q7, s13.6

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.564*
Caret GLM	0.768
Bayesian GLM	1.168

4. **Final features and log-loss table for Forward Selection Algorithm -** q2_all.words, q4_2.attractive, q4_6.cheap, s13.6, q4_5.casual, q4_11.for.someone.like.me, s7.involved.in.the.selection.of.the.cosmetic.products, q4_22.sophisticated, q9.how.likely.would.you.be.to.purchase.this.Deodorant, q4_3.bold, q4_19.ordinary, q4_9.elegant, q8.7, s11.marital.status, q8.20

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.579*
Caret GLM	0.767
Bayesian GLM	1.196

5. **Final features and log-loss table for Backward Elimination Algorithm –** q2_all.words, q4_1.artificial.chemical, q4_2.attractive,q4_3.bold,q4_5.casual,q4_6.cheap, q4_8.easy.to.wear, q4_9.elegant, q4_11.for.someone.like.me, q4_12.heavy, q4_13.high.quality, q4_15.masculine, q4_18.old.fashioned, q4_19.ordinary, q4_20.overpowering, q4_21.sharp, q4_22.sophisticated, q4_23.upscale, q5_1.Deodorant.is.addictive, q7, q8.7, q8.20,q9.how.likely.would.you.be.to.purchase.this.Deodorant,q10.prefer.this.Deodorant.or.your.usual.Deodorant,Q13_Liking.after.30.minutes, q14.Deodorant.overall.

on.a.scale.from.1.to.10, ValSegb, s7.involved.in.the.selection.of.the.cosmetic.products, s10.income, s11.marital.status, s13.6, s13b.bottles.of.Deodorant.do.you.currently.own

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.577*
Caret GLM	0.755
Bayesian GLM	1.208

6. **Final features and log-loss table for Random Forest Feature Importance Algorithm -** q4_6.cheap, q4_18.old.fashioned, q4_8.easy.to.wear, q4_3.bold, q4_19.ordinary, q4_7.clean,q4_23.upscale,q4_22.sophisticated,q4_1.artificial.chemical, q4_2.attractive, q4_21.sharp,q4_12.heavy,q4_11.for.someone.like.me, q10.prefer.this.Deodorant.or.your.usual.Deodorant, q4_20.overpowering, q4_17.natural, q2_all.words, q4_9.elegant, q4_5.casual, s13b.bottles.of.Deodorant.do.you.currently.own,ValSegb, q7, Q13_Liking.after.30.minutes,s10.income,q14.Deodorant.overall.on.a.scale.from.1.to.10

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.567*
Caret GLM	0.771
Bayesian GLM	1.183

Deodorant G -

1. **Top 10 features, Q7 –** q7_1 airy, q7_6 Creamy, q7_11 floral, q7_13 fruity, q7_14 herbal, q7_15 juicy, q7_18 powdery, q7_21 spicy, q7_23 pleasantly sweet, q7_28 woody
2. **Top 10 features, Q8 -** q8.2, q8.3, q8.4, q8.8, q8.9, q8.10, q8.14, q8.15, q8.16, q8.18
3. **Final features and log-loss table for Boruta Feature Selection Algorithm -** q2_all.words, q3_1.strength.of.the.Deodorant,q5_1.Deodorant.is.addictive,q7,q9.how.likely.would.you.be.to.purchase.this.Deodorant, q10.prefer.this.Deodorant.or.your.usual.Deodorant, Q13_Liking.after.30.minutes,q14.Deodorant.overall.on.a.scale.from.1.to.10, q4_2.attractive, q4_10.feminine, q4_22.sophisticated, q7, q8.12

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.539*
Caret GLM	0.716
Bayesian GLM	1.144

4. **Final features and log-loss table for Forward Selection Algorithm -** q5_1.Deodorant.is.addictive,q8.17, s13b.bottles.of.Deodorant.do.you.currently.own, q4_24.well.rounded,q4_10.feminine, q4_5.casual, q4_15.masculine, s8.ethnic.background, Q13_Liking.after.30.minutes, q4_3.bold

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.548*
Caret GLM	0.722
Bayesian GLM	1.167

5. **Final features and log-loss table for Backward Elimination Algorithm –** q2_all.words, q4_1.artificial.chemical, q4_2.attractive,q4_3.bold,q4_5.casual,q4_6.cheap, q4_8.easy.to.wear,q4_9.elegant,q4_10.feminine,q4_12.heavy, q4_13.high.quality,q4_15.masculine,q4_18.old.fashioned, q4_19.ordinary,q4_20.overpowering,q4_21.sharp,q4_23.upscale, q4_24.well.rounded,q5_1.Deodorant.is.addictive,q7,q8.17, q9.how.likely.would.you.be.to.purchase.this.Deodorant,q10.prefer.this.Deodorant.or.your.usual.Deodorant, Q13_Liking.after.30.minutes, q14.Deodorant.overall.on.a.scale.from.1.to.10, ValSegb, s8.ethnic.background, s10.income, s13b.bottles.of.Deodorant.do.you.currently.own

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.548*
Caret GLM	0.741
Bayesian GLM	1.176

6. **Final features and log-loss table for Random Forest Feature Importance Algorithm -** q4_5.casual,q4_10.feminine,q4_11.for.someone.like.me, q4_24.well.rounded, q4_18.old.fashioned,q4_9.elegant, q4_6.cheap,q10.prefer.this.Deodorant.or.your.usual.Deodorant, q4_3.bold,q4_23.upscale,q4_17.natural,q9.how.likely.would.you.be.to.purchase.this.Deodorant, q4_21.sharp,q4_12.heavy,q4_19.ordinary,q4_20.overpowering, q5_1.Deodorant.is.addictive,q4_22.sophisticated,q2_all.words, ValSegb,s13b.bottles.of.Deodorant.do.you.currently.own, q7,Q13_Liking.after.30.minutes,s10.income,q14.Deodorant.overall.on.a.scale.from.1.to.10

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.54*
Caret GLM	0.72
Bayesian GLM	1.151

Deodorant H –

1. **Top 10 features, Q7 –** q7_1 airy, q7_2 ambery, q7_5 cool, q7_11 floral, q7_13 fruity, q7_14 herbal, q7_15 juicy, q7_18 powdery, q7_21 spicy, q7_28 woody
2. **Top 10 features, Q8 -** q8.1, q8.3, q8.4, q8.8, q8.9, q8.10, q8.14, q8.15, q8.16, q8.17
3. **Final features and log-loss table for Boruta Feature Selection Algorithm -** q2_all.words, q3_1.strength.of.the.Deodorant, q5_1.Deodorant.is.addictive, q7, q9.how.likely.would.you.be.to.purchase.this.Deodorant, q10.prefer.this.Deodorant.or.your.usual.Deodorant, Q13_Liking.after.30.minutes, q14.Deodorant.overall.on.a.scale.from.1.to.10, q4_2.attractive, q4_6.cheap, q4_11.for.someone.like.me, q4_17.natural, q4_22.sophisticated, q7, q8.19

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.51*
Caret GLM	0.758
Bayesian GLM	1.171

4. **Final features and log-loss table for Forward Selection Algorithm -** q4_22.sophisticated, Q13_Liking.after.30.minutes, q2_all.words, q8.19, q4_11.for.someone.like.me, q4_6.cheap, q7, q3_1.strength.of.the.Deodorant, q4_4.boring

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.52*
Caret GLM	0.689
Bayesian GLM	1.165

5. **Final features and log-loss table for Backward Elimination Algorithm –** q2_all.words, q3_1.strength.of.the.Deodorant, q4_1.artificial.chemical,q4_2.attractive,q4_3.bold, q4_5.casual,q4_6.cheap,q4_8.easy.to.wear,q4_9.elegant, q4_11.for.someone.like.me, q4_12.heavy,q4_13.high.quality, q4_15.masculine,q4_18.old.fashioned, q4_19.ordinary, q4_20.overpowering,q4_21.sharp,q4_22.sophisticated, q4_23.upscale, q5_1.Deodorant.is.addictive,q7,q8.19, q9.how.likely.would.you.be.to.purchase.this.Deodorant,q10.prefer.this.Deodorant.or.your.usual.Deodorant, Q13_Liking.after.30.minutes, q14.Deodorant.overall.on.a.scale.from.1.to.10, ValSegb, s10.income, s13b.bottles.of.Deodorant.do.you.currently.own

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.518*
Caret GLM	0.699
Bayesian GLM	1.171

6. **Final features and log-loss table for Random Forest Feature Importance Algorithm -** q9.how.likely.would.you.be.to.purchase.this.Deodorant, q4_5.casual, q4_15.masculine, q4_2.attractive, q4_24.well.rounded, q4_9.elegant, q4_6.cheap, s13.8, q4_19.ordinary, q4_21.sharp, q4_1.artificial.chemical,q4_22.sophisticated,q4_11.for.someone.like.me, q4_17.natural,q4_12.heavy,4_18.old.fashioned,q4_20.overpowering, q10.prefer.this.Deodorant.or.your.usual.Deodorant,q2_all.words, s13b.bottles.of.Deodorant.do.you.currently.own,ValSegb, q7,Q13_Liking.after.30.minutes,s10.income,q14.Deodorant.overall.on.a.scale.from.1.to.10

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.507*
Caret GLM	0.789
Bayesian GLM	1.176

Deodorant J –

1. **Top 10 features, Q7 -** q7_1 airy, q7_2 ambery, q7_5 cool, q7_11 floral, q7_13 fruity, q7_14 herbal, q7_15 juicy, q7_18 powdery, q7_23 pleasantly sweet, q7_28 woody
2. **Top 10 features, Q8 -** q8.2, q8.3, q8.4, q8.5, q8.8, q8.9, q8.14, q8.15, q8.16, q8.17
3. **Final features and log-loss table for Boruta Feature Selection Algorithm -** q2_all.words, q3_1.strength.of.the.Deodorant,q5_1.Deodorant.is.addictive,q7,q9.how.likely.would.you.be.to.purchase.this.Deodorant, q10.prefer.this.Deodorant.or.your.usual.Deodorant,

Q13_Liking.after.30.minutes,q14.Deodorant.overall.on.a.scale.from.1.to.10,q4_2.attractive,q4_6.cheap,q4_8.easy.to.wear,q4_9.elegant,q4_11.for.someone.like.me,q4_12.heavy, q4_23.upscale

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.598*
Caret GLM	0.729
Bayesian GLM	1.156

4. **Final features and log-loss table for Forward Selection Algorithm -** q2_all.words, ValSegb,q9.how.likely.would.you.be.to.purchase.this.Deodorant,q3_1.strength.of.the.Deodorant, q4_12.heavy, q8.6, q4_11.for.someone.like.me, q8.5, q8.10, s13b.bottles.of.Deodorant.do.you.currently.own,q8.1,q11.time.of.day.would.this.Deodorant.be.appropriate

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.607*
Caret GLM	0.705
Bayesian GLM	1.153

5. **Final features and log-loss table for Backward Elimination Algorithm –** q2_all.words, q3_1.strength.of.the.Deodorant, q4_1.artificial.chemical,q4_2.attractive,q4_3.bold, q4_5.casual,q4_6.cheap,q4_8.easy.to.wear,q4_9.elegant, q4_12.heavy,q4_13.high.quality,q4_15.masculine,q4_18.old.fashioned, q4_19.ordinary,q4_20.overpowering,q4_21.sharp,q4_23.upscale, q5_1.Deodorant.is.addictive,q7,q8.1,q8.5,q8.6,q8.10, q8.18,q9.how.likely.would.you.be.to.purchase.this.Deodorant, q10.prefer.this.Deodorant.or.your.usual.Deodorant,Q13_Liking.after.30.minutes, q14.Deodorant.overall.on.a.scale.from.1.to.10,ValSegb, s10.income,s13b.bottles.of.Deodorant.do.you.currently.own

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.605*
Caret GLM	0.714
Bayesian GLM	1.193

6. **Final features and log-loss table for Random Forest Feature Importance Algorithm -** q4_3.bold, q4_1.artificial.chemical, q4_15.masculine, q4_5.casual, q4_2.attractive, q9.how.likely.would.you.be.to.purchase.this.Deodorant, q4_6.cheap,q4_13.high.quality, q4_8.easy.to.wear, q4_19.ordinary, q4_21.sharp, q4_9.elegant, q4_23.upscale, q4_12.heavy, q10.prefer.this.Deodorant.or.your.usual.Deodorant,q4_20.overpowering, q5_1.Deodorant.is.addictive,q4_18.old.fashioned,q2_all.words, s13b.bottles.of.Deodorant.do.you.currently.own, q7, ValSegb, Q13_Liking.after.30.minutes, s10.income, q14.Deodorant.overall.on.a.scale.from.1.to.10

Type of Logistic Regression	Logarithmic Loss (Metric)
Simple GLM	0.606*
Caret GLM	0.751
Bayesian GLM	1.167

Conclusion

Across all feature selection algorithms which were applied on this data set these are the primary factors which affect the sales of each deodorant.

Deodorant B –

q2_all.words, q5_1.Deodorant.is.addictive, q7, Q13_Liking.after.30.minutes, q14.Deodorant.overall.on.a.scale.from.1.to.10, q4_13.high.quality, q4_23.upscale

Deodorant F –

q2_all.words, q4_2.attractive, q4_5.casual, q4_9.elegant, q4_11.for.someone.like.me, q4_22.sophisticated

Deodorant G –

q5_1.Deodorant.is.addictive, Q13_Liking.after.30.minutes, q4_10.feminine

Deodorant H –

q2_all.words, q7, Q13_Liking.after.30.minutes, q4_6.cheap, q4_11.for.someone.like.me, q4_22.sophisticated, q7

Deodorant J –

q2_all.words, q9.how.likely.would.you.be.to.purchase.this.Deodorant, q4_12.heavy

ANNEXURE

Annexure 1

Plots for Multiple Correspondence Analysis, used for data transformation or initial variable transformation in Q7 and Q8.

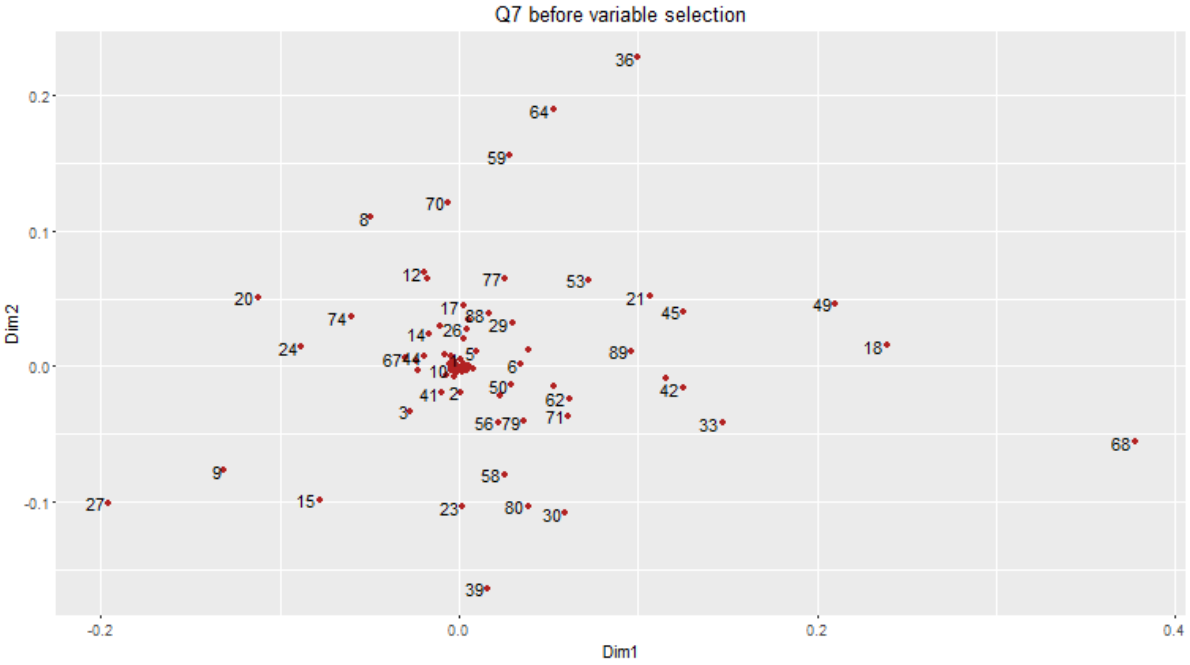

Figure 1 - Multiple Correspondence Analysis graph for Q7 before selecting best variables of Q7

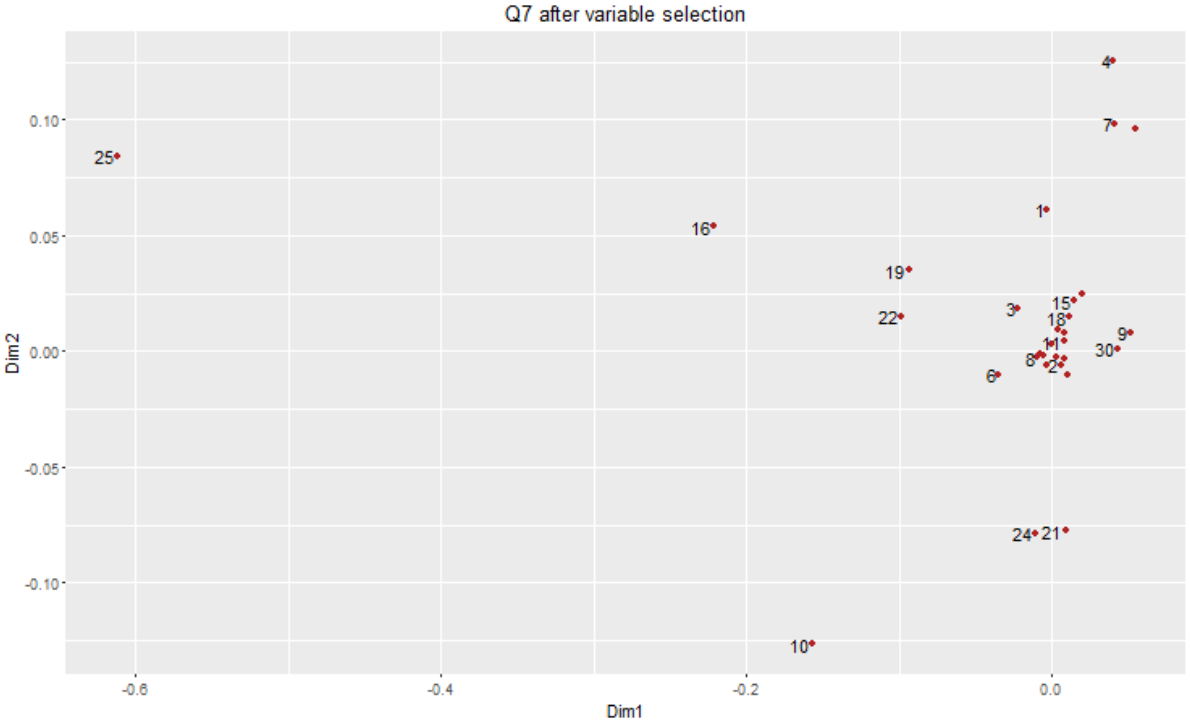

Figure 2 - Multiple Correspondence Analysis graph for Q7 after selecting best variables of Q7

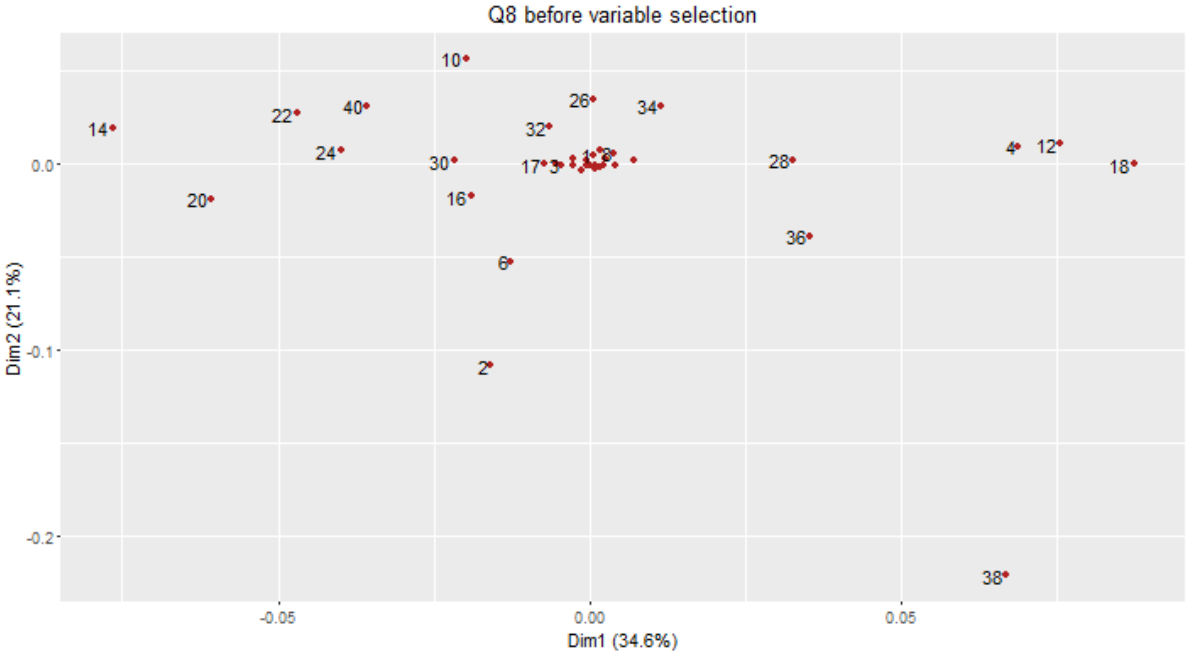

Figure 3 - Multiple Correspondence Analysis graph for Q8 before selecting best variables of Q8

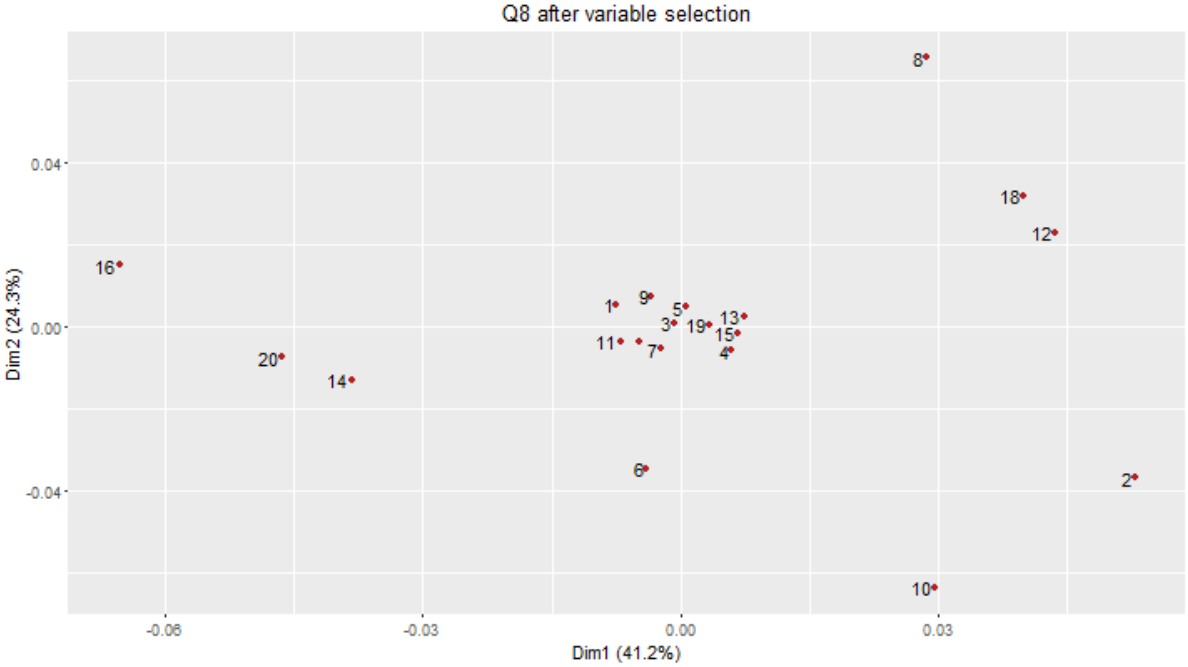

Figure 4 - Multiple Correspondence Analysis graph for Q8 after selecting best variables of Q8

Annexure 2

Coefficient's of best model for each deodorant are:

Deodorant B

Best feature selector = Boruta Feature Selection Algorithm

Best Generalized Linear Model = Simple GLM

Variable Names	Coefficients
(Intercept)	1.11874
q2_all.words	0.12810
q3_1.strength.of.the.Deodorant	0.08446
q5_1.Deodorant.is.addictive	-0.08454
q7	-0.06607
q9.how.likely.would.you.be.to.purchase.this.Deodorant	-0.07475
q10.prefer.this.Deodorant.or.your.usual.Deodorant	0.01101
Q13_Liking.after.30.minutes	-0.04086
q14.Deodorant.overall.on.a.scale.from.1.to.10	-0.03502
q4_2.attractive	-0.03865
q4_8.easy.to.wear	-0.05449
q4_11.for.someone.like.me	-0.03329
q4_13.high.quality	-0.09290
q4_16.memorable	-0.03425
q4_22.sophisticated	-0.03436
q4_23.upscale	-0.10153
q4_24.well.rounded	-0.06521
s13.2	1.01427
s13a.b.most.often	1.18976

Deodorant F

Best feature selector = Boruta Feature Selection Algorithm

Best Generalized Linear Model = Simple GLM

Variable Names	Coefficients
(Intercept)	0.636628
q2_all.words	0.183337
q3_1.strength.of.the.Deodorant	-0.029679
q5_1.Deodorant.is.addictive	0.015587
q7	-0.036736
q9.how.likely.would.you.be.to.purchase.this.Deodorant	0.098695
q10.prefer.this.Deodorant.or.your.usual.Deodorant	0.039780
Q13_Liking.after.30.minutes	-0.022795
q14.Deodorant.overall.on.a.scale.from.1.to.10	0.005748
q4_2.attractive	-0.122889
q4_5.casual	-0.100369
q4_7.clean	-0.057838

q4_9.elegant	-0.092107
q4_11.for.someone.like.me	-0.101494
q4_17.natural	-0.063241
q4_22.sophisticated	-0.094725
s13.6	3.448651

Deodorant G

Best feature selector = Boruta Feature Selection Algorithm

Best Generalized Linear Model = Simple GLM

Variable Names	Coefficients
(Intercept)	0.636628
q2_all.words	0.183337
q3_1.strength.of.the.Deodorant	-0.029679
q5_1.Deodorant.is.addictive	0.015587
q7	-0.036736
q9.how.likely.would.you.be.to.purchase.this.Deodorant	0.098695
q10.prefer.this.Deodorant.or.your.usual.Deodorant	0.039780
Q13_Liking.after.30.minutes	-0.022795
q14.Deodorant.overall.on.a.scale.from.1.to.10	0.005748
q4_2.attractive	-0.122889
q4_5.casual	-0.100369
q4_7.clean	-0.057838
q4_9.elegant	-0.092107
q4_11.for.someone.like.me	-0.101494
q4_17.natural	-0.063241
q4_22.sophisticated	-0.094725
s13.6	3.448651

Deodorant H –

Best feature selector = Random Forest Feature Importance Method

Best Generalized Linear Model = Simple GLM

Variable Names	Coefficients
(Intercept)	1.084568
q9.how.likely.would.you.be.to.purchase.this.Deodorant	0.003669
q4_5.casual	-0.048942
q4_15.masculine	0.012118
q4_2.attractive	-0.032943
q4_24.well.rounded	-0.016832
q4_9.elegant	-0.046298
q4_6.cheap	0.091165
s13.8	0.339338
q4_19.ordinary	-0.062372
q4_21.sharp	-0.012130

q4_1.artificial.chemical	0.025470
q4_22.sophisticated	-0.123650
q4_11.for.someone.like.me	-0.087680
q4_17.natural	-0.032456
q4_12.heavy	0.022599
q4_18.old.fashioned	0.035385
q4_20.overpowering	-0.033041
q10.prefer.this.Deodorant.or.your.usual.Deodorant	-0.025251
q2_all.words	0.110706
s13b.bottles.of.Deodorant.do.you.currently.own	-0.014169
ValSegb	-0.036728
q7	-0.089210
Q13_Liking.after.30.minutes	-0.083033
s10.income	-0.033643
q14.Deodorant.overall.on.a.scale.from.1.to.10	-0.013834

Deodorant J

Best feature selector = Boruta Feature Selection Algorithm

Best Generalized Linear Model = Simple GLM

Variable Names	Coefficients
(Intercept)	-0.5159086
q2_all.words	0.1144397
q3_1.strength.of.the.Deodorant	0.1229265
q5_1.Deodorant.is.addictive	-0.0320782
q7	-0.0900251
q9.how.likely.would.you.be.to.purchase.this.Deodorant	0.0065440
q10.prefer.this.Deodorant.or.your.usual.Deodorant	-0.0350913
Q13_Liking.after.30.minutes	-0.0781145
q14.Deodorant.overall.on.a.scale.from.1.to.10	-0.0131435
q4_2.attractive	-0.0476214
q4_6.cheap	0.0906292
q4_8.easy.to.wear	-0.0005263
q4_9.elegant	-0.0660780
q4_11.for.someone.like.me	-0.0990296
q4_12.heavy	0.0241774
q4_23.upscale	0.0643571