Corelation on relegion:

roman catholic and religion catholic are highly corelated. hence we can remove religion catholic.

religion protestent is highly corelated to United Church and Anglican. Hence we can keep religion protestent and remove United Church and Anglican variables.

corelation on income:

"$20,000 - $34,999, family income" are strongly corelated to "$20,000 - $34,999, household income".hence we will keep "$20,000 - $34,999, household income" and remove "$20,000 - $34,999, family income"

"Under $20,000, family income" is strongly corelated to "Under $20,000, household income".hence we will keep "Under $20,000, household income" and remove "Under $20,000, family income".

"$50,000 and over, household income" , "$50,000 and over, family income","Average income, household income $","Average income, family income $","Median income, family income $" and "Median income, household income $" are strongly corelated

among themselves .Hence we will keep "Average income, household income $" and remove the rest of them.

Corelation on ethinicity AND LANGUAGE:

"amtfrench" and "French origins, ethnic origin" are highly corelated . Hence we will keep "French origins, ethnic origin" and remove "amtfrench".

"amtmultlin" and "amtnengnon" are highly corelated. Hence we will keep "amtmultlin" and remove "amtnengnon".

"British origins, ethnic origin" and "English, single ethnic origin" are highly corelated. Hence we will keep "British origins, ethnic origin" and remove "English, single ethnic origin".

Corelation between family types:

"Female parent" and "Total lone-parent families" are highly corelated. Hence we will keep "Total lone-parent families" and remove "Female parent".

"Total families of now-married couples" and "Total husband-wife families" are highly corelated. Hence we will keep "Total families of now-married couples" and remove "Total husband-wife families".

"1 son or daughter" and "Total lone-parent families" are highly corelated. Hence we will keep "Total lone-parent families" and remove "1 son or daughter".

lgTransSpend—log(Spend per transaction)

Reason: When spend / transaction was calculated, the distribution of the result was skewed. The characteristic of the response is better displayed when the log transformation is performed.

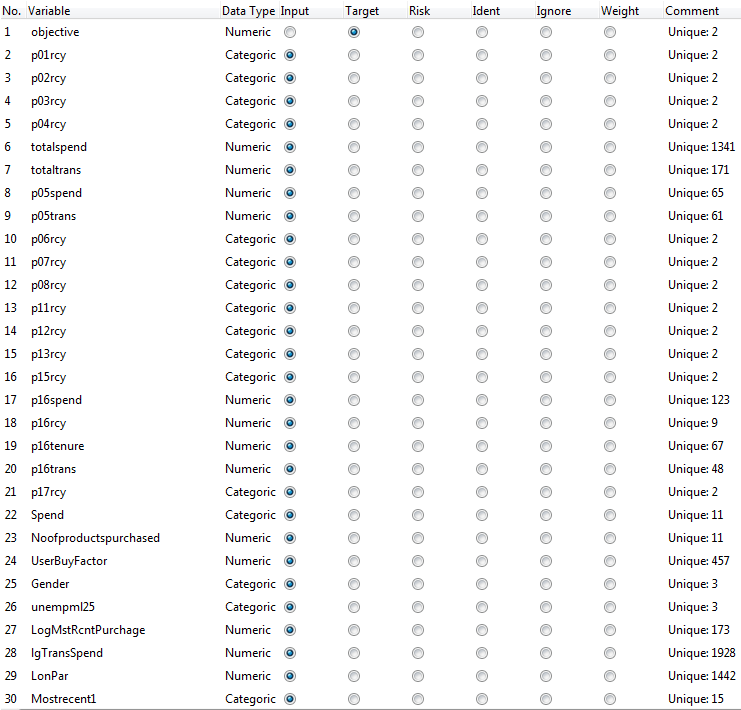
LogMstRcntPurchage- Log(Most Recent purchase)

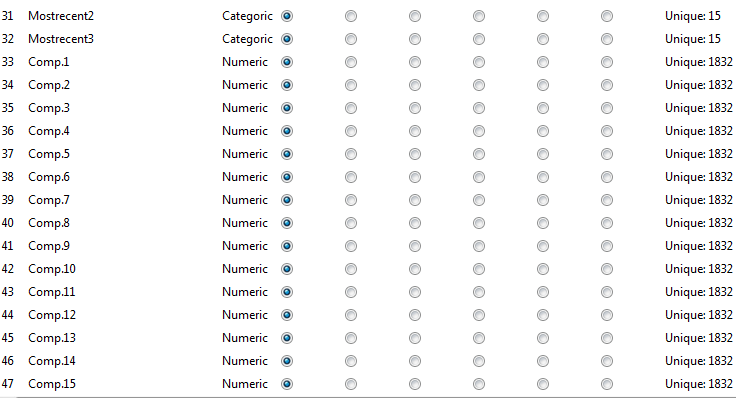
Most recent purchase variable indicates most recent purchase made by the customer, log transformation is performed to reduce the skewness in data and predict the response rate better.

|  |  |  |
| --- | --- | --- |
| Mostrecent1 | Mostrecent2 | Mostrecent3 |

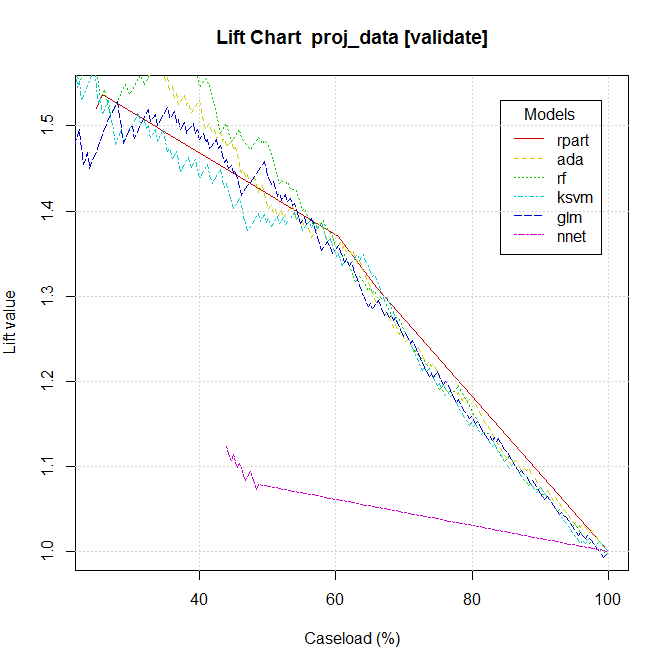
The variables indicate the product based on its recency of purchase, Mostrecent1 indicates the highest most recent purchase

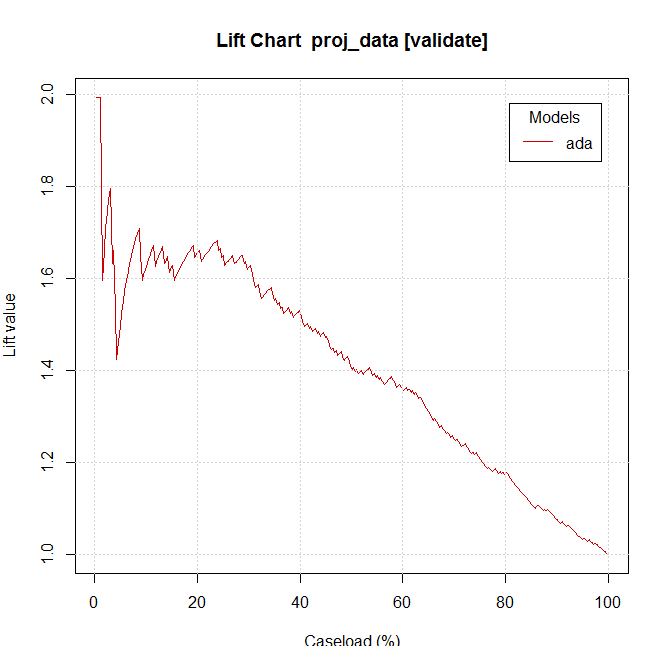
UserBuyFactor – Indicates the tendency of the buyer to purchase the product.

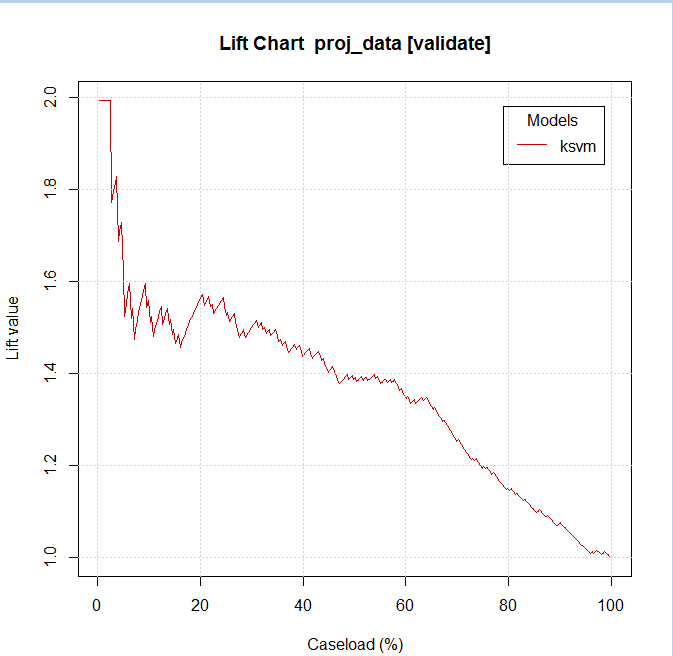


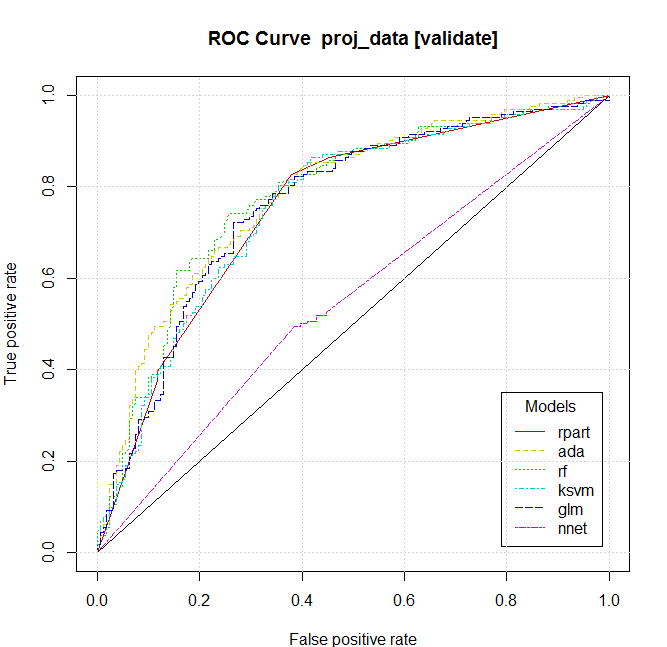


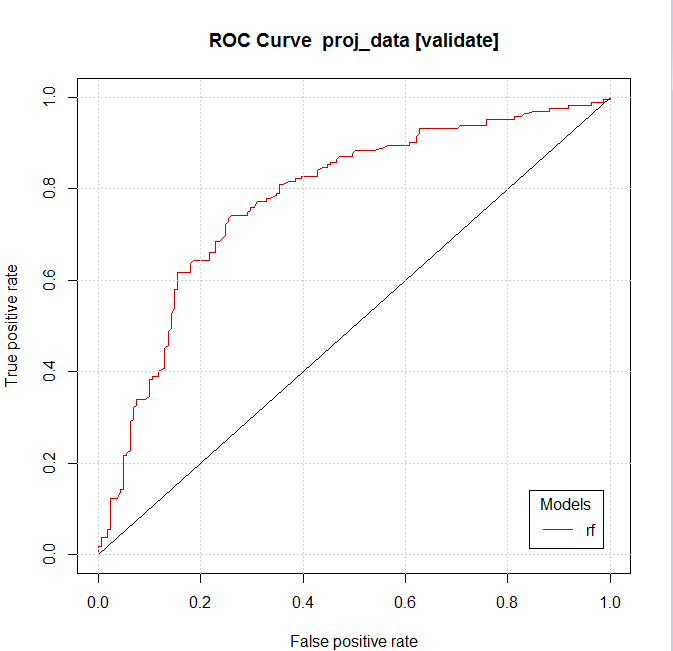
|  |  |  |  |
| --- | --- | --- | --- |
| **Error matrix for the Decision Tree model** | | | |
|  |  |  |  |
| Actual | Predicted | | |
|  | 0 | 1 |
| 0 | 100 | 61 |
| 1 | 28 | 134 |
|  |  |  |  |
| Overall error: 0.2755418 | | |  |
|  |  |  |  |
| **Error matrix for the Ada Boost model** | | | |
|  |  |  |  |
| Actual | Predicted | | |
|  | 0 | 1 |
| 0 | 106 | 55 |
| 1 | 37 | 125 |
|  |  |  |  |
| Overall error: 0.2848297 | | |  |
|  |  |  |  |
| **Error matrix for the Random Forest model** | | | |
|  |  |  |  |
| Actual | Predicted | | |
|  | 0 | 1 |
| 0 | 114 | 47 |
| 1 | 42 | 120 |
|  |  |  |  |
| Overall error: 0.2755418 | | |  |
|  |  |  |  |
| **Error matrix for the SVM model** | | | |
|  |  |  |  |
| Actual | Predicted | | |
|  | 0 | 1 |
| 0 | 107 | 54 |
| 1 | 39 | 123 |
|  |  |  |  |
| Overall error: 0.2879257 | | |  |
|  |  |  |  |
| **Error matrix for the Linear model** | | | |
|  |  |  |  |
| Actual | Predicted | | |
|  | 0 | 1 |
| 0 | 112 | 49 |
| 1 | 43 | 119 |
|  |  |  |  |
| Overall error: 0.2848297 | | |  |
|  |  |  |  |
| **Error matrix for the Neural Net model** | | | |
|  |  |  |  |
| Actual | Predicted | | |
|  | 0 | 1 |
| 0 | 95 | 66 |
| 1 | 81 | 81 |
|  |  |  |  |
| Overall error: 0.4551084 | | |  |











Create an ensemble method and evaluate it.

With error rate, compute a dollar value of profit and loss for all the models.

Derived variable histogram for visualisation.

Modelling, Evaluation --- fine

Link analysis

Clustering