

Question 1

Developing a Mental Model

Three factors we believe to be likely to influence whether a non-RRSP holding VanCity member will acquire a new RRSP term are: (1) a member's ability and need to purchase an RRSP, (2) a member's feelings about and relationship with VanCity, and (3) how often a member interacts with VanCity and is exposed to VanCity products.

The first factor includes features like income, saving, age, tax bracket or investment experience. Potential customers may have tax deferral needs due to higher income. Sufficient cash surplus is also an important prerequisite for making contributions. The second factor explains why clients would purchase RRSPs through VanCity rather than other competitors. It relates to client relationships, how long a relationship is, and whether VanCity is their primary bank, etc. The third factor holds that higher interaction frequency leads to higher exposure, increasing the probability for members to indirectly gain product knowledge, thereby stimulating spontaneous purchases. A high interaction rate may also suggest that customers are more likely to receive leads from VanCity campaigns or sales teams, increasing purchase propensity.

Examining the Variables in the Dataset and Data Cleaning

To get started, we load in R all necessary packages, source the helper functions, and read in the data file. We name it *rrsp* and treat all empty strings as *NAs* using `na.strings = ""` in `read.csv()`. The raw data contains 5110 records and has been oversampled with a 50/50 split of RRSP purchasers and non-purchasers. We use `create.samples(rrsp, est = 0.5, val = 0.5, rand.seed = 1)` to create 50/50 estimation and validation samples with no holdout sample.

Next, we examine if there are variables related to the three factors of our mental model. We find age, total monthly deposits, balances of different accounts, and new acquisitions of loans and mortgages to be potentially good measures of the first factor. Additionally, the average income and investment income at the community level are also good geo-demographic proxies of a member's purchase ability and need. Whether a member owns an ATM card or sets up a payroll deposit may reflect his/her feelings and relationship with VanCity. The frequency of in-branch transactions, the average total monthly deposits, the total number of distinct services, and changes in the number of services and products are also potential measures related to feelings and relationships. The third factor may be measured by the balances of different accounts, the frequency of transactions across different channels, the number of distinct services held, or the number of contributors and contributions in the neighborhood a member resides in. With all these in mind, we start to clean data. In Figure 1, we examine the summary of variables using `variable.summary()`.

Look for potential problems with missing values

There are 7 particularly problematic variables with a missing rate greater than 3% including *numrr_1*, *numcon_1*, *BALCHQ*, *BALSAV*, *BALLOC*, *BALMRGG*, and *BALLOAN*. Less problematic ones are *CH_NM_SERV*, *CH_NM_PRD*, *valsegm*, *pcode*, *avginc_1*, *avginv_1*, and *N_IND_INC_*.

We first investigate why *numrr_1* and *numcon_1* have about 8.5% values missing. 70 records have missing postal codes, resulting in disjointed data at the individual and community levels. Except for those records, we believe that they are most likely meaningful missings. The range of *numrr_1* has the minimal value is 20 instead of 0, so we assume that *NAs* may represent communities without RRSP holders (and thus no contribution). Before we handle these meaningful *NAs* by imputing zero using `if_else(is.na(rrsp$numcon_1), 0L, rrsp$numcon_1)`, there are two steps necessary. First, since individuals with missing zip codes may live in communities with arbitrary levels of *numrr_1* and *numcon_1*, to avoid zero imputation for such cases, we remove the 70 records using `rrsp <- rrsp[!is.na(rrsp$pcode),]`. Second, 5 records are in a community with 160 contributors and 0 contributions. An option to distinguish these zeros from the imputed zeros is to add a dummy variable. However, since there are only 5 such observations, we did not do it to prevent a worthless and zero inflated indicator. The rest of the particularly problematic variables are the 5 balance variables. Though the *NA%* for these variables are extremely high, we do not have a strong reason to remove any of them. By checking the ranges, we see that the minimal amount of *BALCHQ*, *BALSAV*, and *BALLOC* is 0, meaning that they have an account of the type but with a balance of 0. For those *NAs*, it's reasonable to think of them as “no balance”, representing members without that type of account. Similarly, a mortgage or loan with 0 balance does not make sense (i.e., such accounts should have been closed). So we assume that the *NAs* in *BALLOAN* and *BALMRGG* are meaningful missings: these individuals do not have a loan or mortgage. Before we impute zeros for the 5 balance variables, an important step is to add dummy indicators to distinguish having an account or not, considering the nature of these features and the size of accounts with a 0 balance. As an example, we use `rrsp$BALCHQ_IND <- if_else(is.na(rrsp$BALCHQ), 1, 0)` to add an indicator and `rrsp$BALCHQ <- if_else(is.na(rrsp$BALCHQ), 0, rrsp$BALCHQ)` to impute zero. For the 7 less problematic variables, we will explore them later.

Thin factor level

The *pcode* variable has a thin factor level of 1 and is statistically irrelevant, although we may use this variable for profiling. We run `rrsp$pcode <- NULL` to remove the column for now. Another variable with the similar issue is *valsegm*. By running `rrsp %>% group_by(valsegm) %>% summarise(num_row = length(valsegm))`, we see that the only level containing a small number of records is the *NA* level. We will use `na.omit()` to remove the level later.

Unary variables

We do not have unary variables in this dataset.

Trivially related variable problems

We first examine if the variable *valsegm* could potentially be a trivially related variable. It depends on whether an RRSP purchase would increase the rank or not. We do not have this information yet and this possibility may be determined in the subsequent analysis or in further

discussions with VanCity. We decided to keep them for now. In fact, our analysis shows that *valsegm* is likely not a trivially related variable.

CH_NM_SERV and *CH_NM_PRD* are trivially related variables because RRSP purchases are likely to “cause” changes in these two predictors. We remove them using `rrsp$CH_NM_PRD <- NULL` and `rrsp$CH_NM_SERV <- NULL`.

For the variable *TOTSERV*, it is more complicated than the previous three features. The variable may cause target leakage because it is most likely measured after an individual has purchased an RRSP term, resulting in an increase of 1 in *TOTSERV*. Another concern is whether members tend to obtain more than 1 service because of an RRSP purchase (e.g., a minimum number of distinct products to waive monthly fees). However, based on our mental map, we believe that *TOTSERV* should be a decent predictor as it strongly reflects customer intimacy, exposure, product knowledge, financial situation, investment experience and more. That is, we believe that the more diverse the client’s product category is, it should be more likely for the client to make an RRSP purchase. Additionally, since our objective is to cross-sell to “existing” clients, it's very likely that they have at least 1 product line. So, it's likely that an increase of 1 in *TOTSERV* due to the RRSP purchase would not make *TOTSERV* a bad predictor.

With these considerations, we decide to test if, with some transformation, we can turn *TOTSERV* into a powerful predictor and prevent the possible target leakage. To do this, we assume that *TOTSERV* increases by at most 1 for each RRSP purchase. We form a new feature called *TOTSERV.NEW* to hold the values of “the total number of services and/or product lines an individual has BEFORE the purchase of an RRSP”. The values are obtained by the following code: `rrsp$TOTSERV.NEW <- if_else(rrsp$APURCH == "Y", if_else(rrsp$TOTSERV < 1, 0, rrsp$TOTSERV - 1), rrsp$TOTSERV)`.

Database housekeeping variable problems

The variable *unique* is the record keeping identification number of each person in the database. We set it as a row name using `row.names(rrsp) <- rrsp$unique` and remove it by `rrsp$unique <- NULL`. The gender variable takes two columns. We remove one of them using `rrsp$gendf <- NULL` so that female is the base level. The variable *N_IND_INC_* appears to be statistically irrelevant. We remove it using `rrsp$N_IND_INC_ <- NULL`. Ten binary variables should be of a factor class instead of numeric. Though optional, we factorize them using `rrsp$paydep <- factor(rrsp$paydep)`.

Final check

We use `sapply(select_if(rrsp,is.numeric),summary)` and find two cases with age 0. We remove the two records using `rrsp <- rrsp[rrsp$age != 0,]`. We run `variable.summary()` to conduct a final check. As shown in Figure 2, through the previous steps of data cleaning, we have only 3 variables with a missing rate of 0.26 ~ 1.09%, including the *valsegm* variable with the thin factor level problem. Removing these records does not lose as much good information because they have NAs across many variables. Using `na.omit()`, we lose another 68 observations, or about

1.35% of 5038 observations. In total, we removed 70 (with missing *pcode*) + 68 (with missing *valsegm*, *avginc_1*, and *avginv_1*) + 2 (aged 0) = 140 cases from the original dataset. or about 2.74% of 5110 observations, which is acceptable. The cleaned data is stored in *rrsp2*.

Modelling the Data

Test if TOTSERV.NEW is a trivially related variable

We built a random forest using all variables on the estimation sample with 500 trees and 4 variables per try. The modeling code and its confusion table is in Figure 3. The table shows that the misclassification rate from the internal “out of bag” validation set is 0.22 for “N” and 0.18 for “Y”. The overall hit rate is 79.67%, which is a decent improvement over the 50% hit rate we would get without a model. The importance plot in Figure 4 shows some potential issues with how extremely predictive the *TOTSERV.NEW* variable is.

For the *APURCH* target variable, we confirm that the level being predicted is “Y” using `levels(rrsp2$APURCH)`. We use the `partial()` function to create multiple partial dependence plots (PDP’s) before and after trimming the top 10% to visualize the relationships between some important features and the purchase behavior. Notably, *TOTSERV.NEW* is decreasing at a decreasing rate (Figure 5). Additionally, the PDP of *valsegm* (Figure 6) shows that Group E has the lowest effect and can be used as a base level by `rrsp2$valsegm<-relevel(rrsp2$valsegm, "E")`. Before building a regression model, we plot the correlation matrix (Figure 7) of numeric predictors using `corrMatrix <- cor(select_if(rrsp2, is.numeric))` and `corrplot(corrMatrix, method = "number", type="lower", diag = FALSE, number.cex = 0.5)`. The plot shows that, except for the transformed *TOTSERV.NEW* and its original *TOTSERV*, only 2 pairs of variables are highly correlated: *numcon_1* and *numrr_1* (0.99), and *avginv_1* and *avginc_1* (0.71). We expect that the logistic model would not identify both *numcon_1* and *numrr_1* as highly significant whereas they are ranked 8th and 10th in the random forest. Similarly, *avginc_1* and *avginv_1* would not appear both significant in the regression while they are ranked 5th and 6th in the forest.

Now we are ready to run the maximal logistic model. The code and the results are shown in Figure 8. The model obtains an AIC of 1586.6 and a McFadden R-squared of 0.559. Before running a stepwise logistic regression, we check the target variable level for the logistic model. Fortunately, since “N” precedes “Y” alphabetically, the base level 1 is “N” and thus the probabilities for level 2 “Y” are calculated, and we can interpret the signs of the coefficients relative to the probability of purchasing an RRSP. Using the maximal logistic model, we run `rrspStep <- step(rrspLogis, direction = "both")` to build a stepwise model, whose AIC score is 1568.6 and McFadden R-squared 0.557. However, when we use a cumulative lift chart on the validation sample with a true response rate of 0.022, we find an extraordinary fit of the validation set (Figure 9). We conclude that *TOTSERV* is indeed a trivially related variable. Therefore, we remove the variable from the forest and the regression model.

Modeling the data without TOTSERV

We run the same forest model without the *TOTSERV* predictor and develop a new hit rate matrix (Figure 10). The hit rate drops from 79% to 63%, showing only 13% improvement over the 50% hit rate we would get if we had to randomly target individuals in a 50/50 database. The new importance plot in Figure 11 has a more normal appearance. The top 5 important predictors are *TOTDEP*, *age*, *BALCHQ*, *avginc_1*, and *avginv_1*. By running the same logistic model without *TOTSERV*, the AIC significantly increases from 1568.6 to 3260.2 and the McFadden R-squared drops from 0.559 to 0.074 (Figure 12). The new stepwise regression model improves the two statistics to 3238.8 and 0.071 (Figure 13). By comparing the performance of the stepwise model and the maximal logistic model on the validation sample in a lift chart (Figure 14), we see that the subset of variables selected by the stepwise algorithm (from 28 to 13) increases both the interpretability and the predictability. However, if we compare the two regression models with the random forest on the validation sample (Figure 15), the forest is performing much better than the stepwise model, indicating that the forest is capturing some non-linear relationships that the logistic models cannot. Therefore, to improve our regression models, we need to check multicollinearity issues and transform variables that have non-linear relationships with the purchase propensity.

Look for improvements

Since interpretation is important, we need to improve our regression model. Figure 16 compares the top 15 important variables ranked by the regression (by significance) and the forest model. The table shows that only 7 variables (highlighted in yellow) are shared in both models. To explain why more than half of the important predictors are eliminated, we identified 3 major possible improvements.

- 1) *TOTDEP* is the most important feature in the forest but has a low significance level and a low magnitude of coefficient. It is not correlated with other variables. However, the 10% trimmed PDP of *TOTDEP* (Figure 17) shows strong concavity. Therefore, we need to log-transform the variable to capture the non-linearity. We use `range(rrsp2$TOTDEP)` and find the minimum value to be 0. So we will use `rrsp2$Log.TOTDEP <- log(rrsp2$TOTDEP + 1)` to avoid the log(0) problem.
- 2) Referring to the 2 highly correlated pairs of variables we have identified, there are multicollinearity issues. Both *numcon_1* and *numrr_1* ranked 7th and 8th in random forest are eliminated by the stepwise model because they are highly correlated (0.99). Similarly, because there is a strong correlation (0.71) between *avginc_1* and *avginv_1*, the average income becomes insignificant in the stepwise model. To handle the issues, we decide to perform a principal component analysis (PCA) and transform each pair into one principal component. The predictors in each pair have different scales. Therefore, we need to normalize them by setting `scales. = T` in `PC.numcon_numrr <- prcomp (select (rrsp2, numcon_1, numrr_1), scale. = T)`. The analysis shows that, *numcon_1* and *numrr_1* can be combined into one variable which explains 99.71% of the variance. Therefore, we

added a new feature to replace *numcon_1* and *numrr_1* by running `rrsp2$PC.numcon.numrr <- PC.numcon_numrr$x[,1]`. Similarly, *avginc_1* and *avginv_1* can be combined into one principal component that explains 85.63% of the variance. We will call the new predictor *PC.avginc.avginv* and add it to the data frame.

- 3) There are other numeric variables that may have a nonlinear relationship with the target variable. We are mainly interested in the 5 balance predictors and the 6 transaction predictors. We will log-transform all of them and select a good subset by trial and error. By `sapply()` a `summary()` function on the predictors, we see that all of them have a minimum value of 0. Therefore, we will add 1 before logging.

We build 3 sets of models where each set has a maximal logistic model and its stepwise model. The first set called *Mixed1* tests the first and the second possible improvements. The second set *Mixed2* and the third set *Mixed3* examine the effect of the log transformation on the 5 balance variables and the 6 transaction variables, respectively. The fit of the stepwise model in each set is steadily improving (i.e., lower AIC and higher McFadden R-squared) as we include more logarithm and PC transformations.

We are also interested in a fourth set of models for further improvements by determining whether a variable should be transformed. In Figure 18, we compare the absolute values of $\Pr(>|z|)$ for all the transformed predictors in the maximal model in each set. We find that the only 3 variables which the log transformation did not improve are *BALLOAN*, *BALMRGG*, and *TXATM*. Therefore, in the fourth set of models, we will use the original form of these 3 features. Additionally, the 2 PCs aimed to eliminate the multicollinearity issues are not significant in all the maximal models. In fact, the principal components are rejected by all the stepwise regression models. Nevertheless, we will include them in the last set of models and later explore why they are important in the forest but not in the stepwise model. Figures 19 and 20 show how we build the *Mixed4* maximal logistic model and its stepwise model.

Comparing Results Between Models and Model Selection

Check fit statistics

		<i>Mixed1</i>	<i>Mixed2</i>	<i>Mixed3</i>	<i>Mixed4</i>
Maximal	AIC	3189.3	3185.1	3181.5	3181.3
	MR2	0.094	0.095	0.096	0.096
Stepwise	AIC	3163.8	3161.5	3160.2	3159.2
	MR2	0.088	0.093	0.094	0.094

The above table compares the fit statistics of all four sets of models. Horizontally, the fit of models is improving. The *Mixed4* stepwise model is the best model with a minimum AIC value of 3159.2 and a high McFadden R-squared of 0.094.

Check cumulative lift charts on the validation sample

All the lift charts will use `trueResp = 0.022` in their `lift.chart()` functions. Figure 21 compares the performance of all 4 maximal models with the initial maximal logistic model. The maximal model of *Mixed2*, *Mixed3*, and *Mixed4* have a similar performance. Figure 22 compares the performance of all 4 stepwise models with the initial stepwise model. The stepwise model of *Mixed3* and *Mixed4* have similar performance as their lifts almost overlap each other. Figure 23 compares the performance of the best two stepwise models with the random forest. The lift chart shows that we have improved the performance of our stepwise model to just as good as that of the random forest. Though the stepwise models of *Mixed3* and *Mixed4* have similar lifts, we choose *Mixed4* for its better fit statistics.

Check neural network models

We build two neural network models as a final check for nonlinearities. The first model is a 4-node single-layer neural network using only the variables chosen by the *Mixed4* stepwise model and a weight decay of 0.15. Figure 24 shows that the neural network performs worse than both the stepwise regression and the forest, meaning that there are likely no detectable patterns in the relations between the predictors that the stepwise regression used and the target but did not capture. Keeping all other parameters the same, the second neural network includes all the original predictors to check nonlinearities in the variables that the stepwise model rejected. Figure 25 shows that the addition of variables to the neural network results in a worse performance, indicating that the extra variables are causing an overfitting issue.

Select the best model

The random forest model and the *Mixed4* stepwise model appear to perform equally well on the validation sample. Additionally, the *Mixed4* stepwise model has a higher lift and less selected predictors than Mr. Lo's model has. Because interpretation is important, the stepwise model is the best choice as it is a good predicting model and is clearly interpretable. The random forest on the other hand provides additional interpretation information on strongly predictive variables that the stepwise regression rejected. In this case, we need to examine why geo-demographic variables are rejected by the stepwise model.

For the *avginc_1* and *avginv_1* variables (and their PC), they may not be representative enough to reflect true income at the individual level. Therefore, we cannot find correlations between them and other predictors (such as *TOTDEP*) that contain similar or even better information about individual income. Since purchasing behavior is expected to be strongly correlated with individual income, *avginc_1* and *avginv_1* may be replaced with seemingly uncorrelated variables that are better predictors of individual-level income. The *numrr_1* and *numcon_1* (and their PCs) variables may be rejected for a similar reason. We speculate that living in a neighborhood with a high RRSP ownership rate may not, by itself, affect purchasing behavior. When there are other variables that contain similar or even better information at the individual level, they may jointly replace *numrr_1* and *numcon_1*.

Question 1 Appendix

```
> ##### Check variable summary #####
> variable.summary(rrsp)
```

	Class	%NA	Levels	Min.Level.Size	Mean	SD
APURCH	factor	0.0000000	2	2555	NA	NA
unique	integer	0.0000000	NA	NA	1.330403e+04	7.651237e+03
age	integer	0.0000000	NA	NA	4.379511e+01	1.190954e+01
gendf	integer	0.0000000	NA	NA	4.346380e-01	4.957579e-01
gendm	integer	0.0000000	NA	NA	5.553816e-01	4.969720e-01
atmcrd	integer	0.0000000	NA	NA	8.045010e-01	3.966232e-01
paydep	integer	0.0000000	NA	NA	3.221135e-01	4.673319e-01
TOTDEP	numeric	0.0000000	NA	NA	9.353147e+03	2.678552e+04
NEWLOC	integer	0.0000000	NA	NA	2.935421e-02	1.688138e-01
NEWMRGG	integer	0.0000000	NA	NA	2.387476e-02	1.526739e-01
TXBRAN	numeric	0.0000000	NA	NA	2.020787e+00	2.765819e+00
TXATM	numeric	0.0000000	NA	NA	3.618546e+00	4.979280e+00
TXPOS	numeric	0.0000000	NA	NA	8.434219e+00	1.421844e+01
TXCHQ	numeric	0.0000000	NA	NA	2.111136e+00	3.210656e+00
TXWEB	numeric	0.0000000	NA	NA	1.272680e+00	2.672658e+00
TXTEL	numeric	0.0000000	NA	NA	7.028702e-03	6.072220e-02
TOTSERV	numeric	0.0000000	NA	NA	3.446519e+00	1.111718e+00
Sample	character	0.0000000	NA	NA	NA	NA
CH_NM_SERV	integer	0.9589041	NA	NA	-9.879470e-03	6.071878e-01
valsegm	factor	1.0763209	6	55	NA	NA
CH_NM_PRD	integer	1.0958904	NA	NA	-6.193114e-02	1.380049e+00
pcode	factor	1.3698630	4548	1	NA	NA
N_IND_INC_	integer	1.6242661	NA	NA	1.631100e+03	3.928495e+03
avginc_1	numeric	1.6242661	NA	NA	3.261148e+04	1.050185e+04
avginv_1	numeric	1.6242661	NA	NA	3.896229e+03	3.172915e+03
numrr_1	integer	8.5322896	NA	NA	6.140287e+02	1.491946e+03
numcon_1	integer	8.5322896	NA	NA	2.802360e+03	6.811798e+03
BALCHQ	numeric	12.0352250	NA	NA	2.811116e+03	6.369533e+03
BALSAV	numeric	38.7084149	NA	NA	1.359478e+03	3.922424e+03
BALLOC	numeric	58.8845401	NA	NA	1.179504e+04	3.783575e+04
BALMRGG	numeric	74.9706458	NA	NA	1.547103e+05	9.630085e+04
BALLOAN	numeric	81.3502935	NA	NA	9.529647e+03	9.307624e+03

Figure 1: Variable Summary (Raw Data)


```
> variable.summary(rrsp) # 5038 observations
```

	Class	%NA	Levels	Min.Level	Size	Mean	SD
APURCH	factor	0.0000000	2		2516	NA	NA
age	integer	0.0000000	NA		NA	4.384736e+01	1.189601e+01
gendm	factor	0.0000000	2		2246	NA	NA
atmcrd	factor	0.0000000	2		983	NA	NA
paydep	factor	0.0000000	2		1626	NA	NA
BALCHQ	numeric	0.0000000	NA		NA	2.489625e+03	6.075110e+03
BALSAV	numeric	0.0000000	NA		NA	8.395877e+02	3.160093e+03
TOTDEP	numeric	0.0000000	NA		NA	9.439346e+03	2.695073e+04
BALLOAN	numeric	0.0000000	NA		NA	1.775293e+03	5.466147e+03
BALLOC	numeric	0.0000000	NA		NA	4.865834e+03	2.507510e+04
BALMRGG	numeric	0.0000000	NA		NA	3.870779e+04	8.209689e+04
NEWLOC	factor	0.0000000	2		150	NA	NA
NEWMRGG	factor	0.0000000	2		120	NA	NA
TXBRAN	numeric	0.0000000	NA		NA	2.030628e+00	2.773389e+00
TXATM	numeric	0.0000000	NA		NA	3.624938e+00	4.982996e+00
TXPOS	numeric	0.0000000	NA		NA	8.465104e+00	1.426287e+01
TXCHQ	numeric	0.0000000	NA		NA	2.117191e+00	3.213903e+00
TXWEB	numeric	0.0000000	NA		NA	1.276147e+00	2.678011e+00
TXTEL	numeric	0.0000000	NA		NA	7.029906e-03	6.107005e-02
TOTSERV	numeric	0.0000000	NA		NA	3.451352e+00	1.109217e+00
numrr_1	integer	0.0000000	NA		NA	5.695256e+02	1.445834e+03
numcon_1	integer	0.0000000	NA		NA	2.599255e+03	6.601230e+03
Sample	character	0.0000000	NA		NA	NA	NA
BALCHQ_IND	factor	0.0000000	2		602	NA	NA
BALSAV_IND	factor	0.0000000	2		1950	NA	NA
BALLOC_IND	factor	0.0000000	2		2075	NA	NA
BALLOAN_IND	factor	0.0000000	2		941	NA	NA
BALMRGG_IND	factor	0.0000000	2		1265	NA	NA
TOTSERV.NEW	numeric	0.0000000	NA		NA	2.951967e+00	1.120121e+00
avginc_1	numeric	0.2580389	NA		NA	3.261423e+04	1.050233e+04
avginv_1	numeric	0.2580389	NA		NA	3.896858e+03	3.173344e+03
valsegm	factor	1.0917031	6		55	NA	NA

Figure 2: Variable Summary (Final Check)

```

255
256 # Main task: test if TOTSERV.NEW is a trivially related variable.
257
258 # [10.3.4] Explore with Forests and Regression
259
260 ##### Random Forest #####
261 # paste(names(rrsp2), collapse = " + ")
262 rrspForestAll <- randomForest(formula = APURCH ~ age + gendm + atmcrd + paydep +
263                               BALCHQ + BALSAV + TOTDEP + BALLOAN + BALLOC +
264                               BALMRGG + NEWLOC + NEWMRGG + TXBRAN + TXATM +
265                               TXPOS + TXCHQ + TXWEB + TXTEL + TOTSERV.NEW +
266                               valsegm + numrr_1 + numcon_1 + avginc_1 + avginv_1 +
267                               BALCHQ_IND + BALSAV_IND + BALLOAN_IND +
268                               BALLOC_IND + BALMRGG_IND,
269                               data = filter(rrsp2, Sample == "Estimation"),
270                               importance = TRUE,
271                               ntree = 500, mtry = 4)
272 # Contingency Table
273 rrspForestAll[["confusion"]]
274 # Misclassification rate from the internal "out of bag" validation set is 0.2229455
275 # for "N" and 0.1841270 for "Y". It means that the overall hit rate is 79.67% or
276 # an error rate of 20.33%. That's a decent improvement over the 50% hit rate we
277 # would get without a model or, equivalently, the 50% response rate we would get
278 # if we had to randomly target individuals in a 50/50 database.
279
280 # Plot the relative importance of the variables:
281 varImpPlot(rrspForestAll, type = 2,
282            main = "rrspForestAll", # title
283            cex = 0.7) # font size
284
287:1 # Random Forest

```

Console Terminal Background Jobs

R 4.1.2 · ~/Desktop/sfu/s13/BUS445/Assignment2/ ↗

```

> ##### Random Forest #####
> # paste(names(rrsp2), collapse = " + ")
> rrspForestAll <- randomForest(formula = APURCH ~ age + gendm + atmcrd + paydep +
+                               BALCHQ + BALSAV + TOTDEP + BALLOAN + BALLOC +
+                               BALMRGG + NEWLOC + NEWMRGG + TXBRAN + TXATM +
+                               TXPOS + TXCHQ + TXWEB + TXTEL + TOTSERV.NEW +
+                               valsegm + numrr_1 + numcon_1 + avginc_1 + avginv_1 +
+                               BALCHQ_IND + BALSAV_IND + BALLOAN_IND +
+                               BALLOC_IND + BALMRGG_IND,
+                               data = filter(rrsp2, Sample == "Estimation"),
+                               importance = TRUE,
+                               ntree = 500, mtry = 4)
> # Contingency Table
> rrspForestAll[["confusion"]]
  N    Y class.error
N 955 274  0.2229455
Y 232 1028 0.1841270

```

Figure 3: Building a Random Forest with *TOTSERV.NEW*

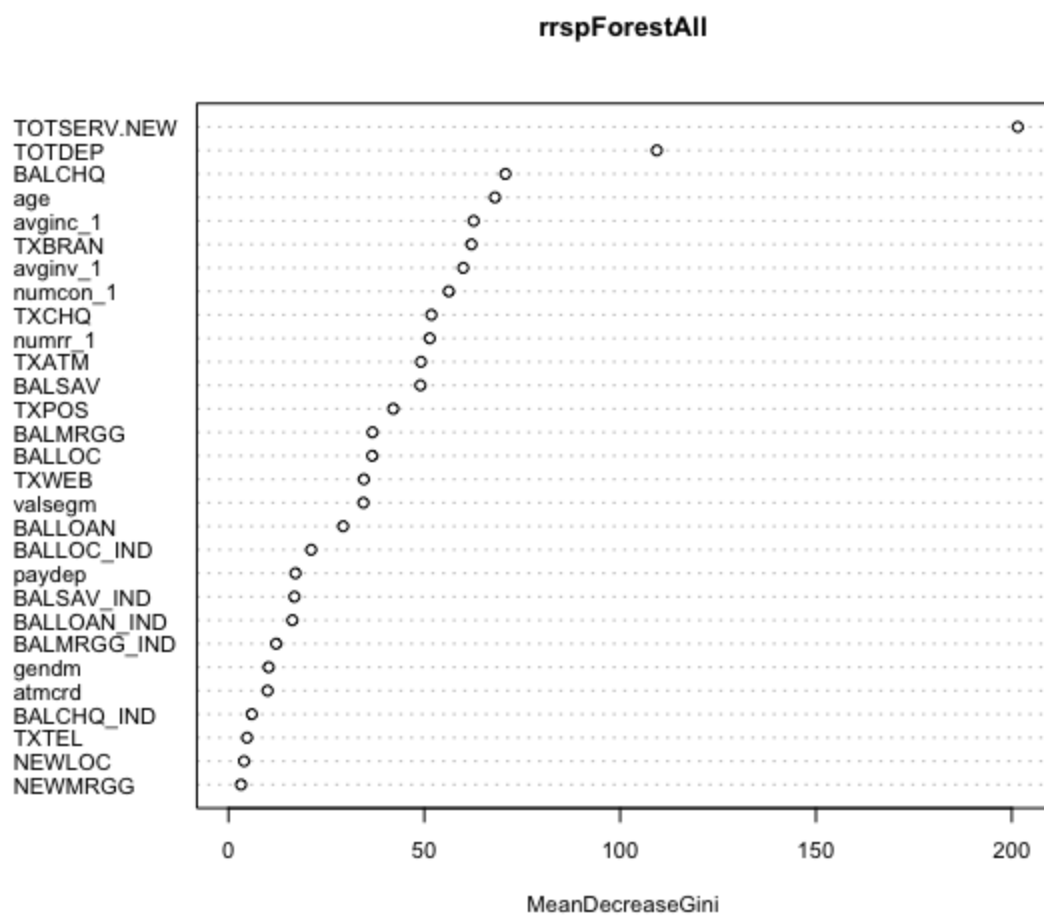


Figure 4: Importance Plot of the First Random Forest model

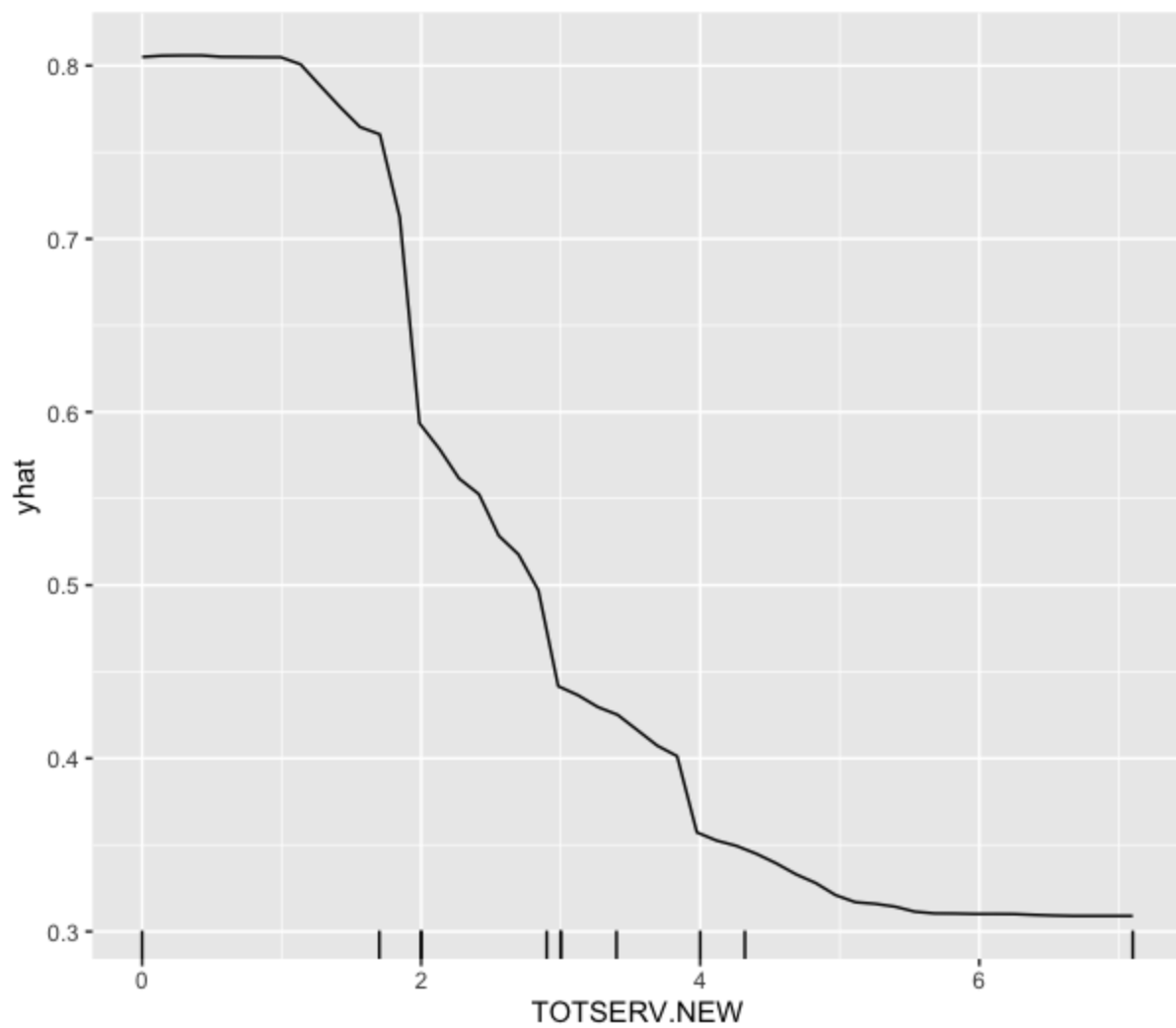


Figure 5: Partial Dependence Plot of *TOTSERV.NEW*

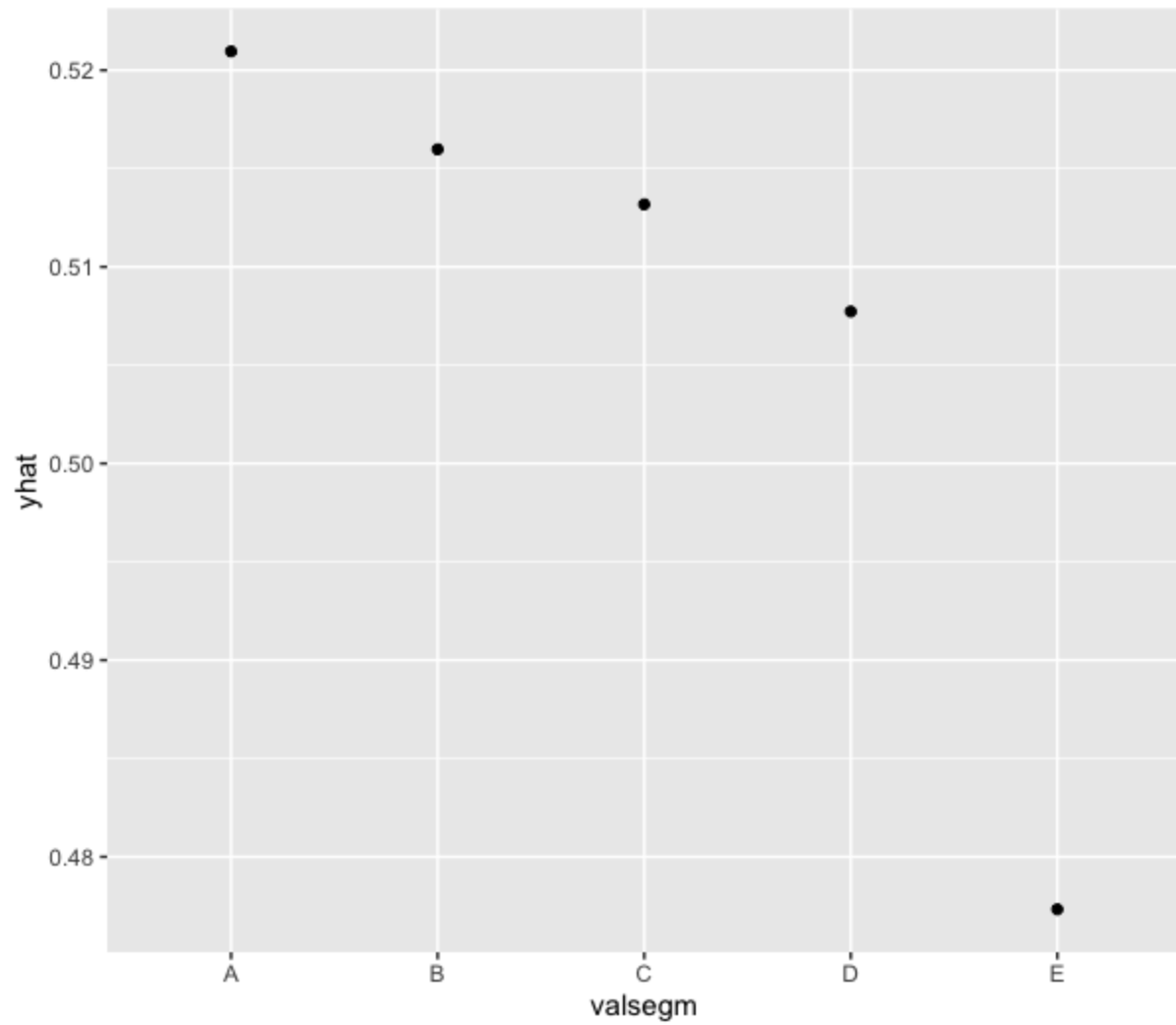


Figure 6: Partial Dependence Plot of *valsegm*

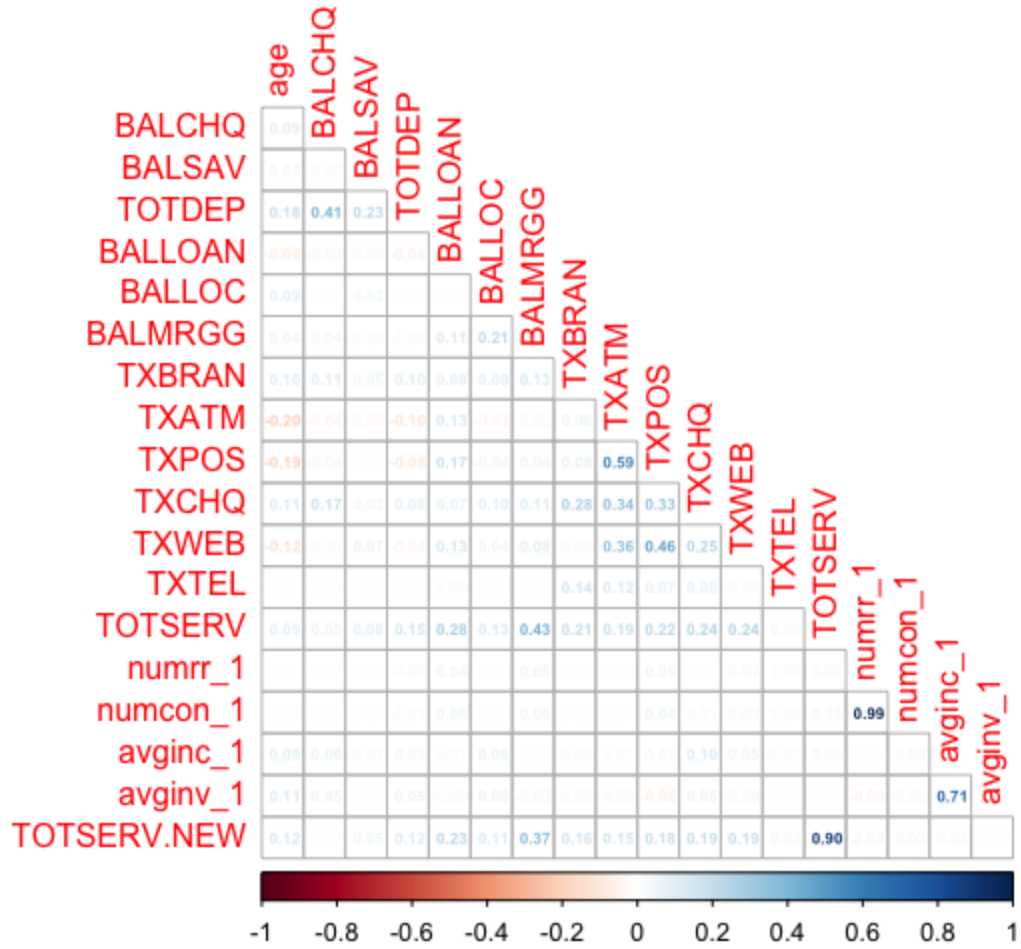


Figure 7: Correlation Plot

```

> # Create a logistic regression model
> rrsplgis <- glm(formula = APURCH ~ age + gendm + atmcrd + paydep + BALCHQ +
+                 BALSAV + TOTDEP + BALLOAN + BALLOC + BALMRGG + NEWLOC +
+                 NEWMRGG + TXBRAN + TXATM + TXPOS + TXCHQ + TXWEB + TXTEL +
+                 TOTSERV.NEW + valsegm + numrr_1 +
+                 numcon_1 + avginc_1 + avginv_1 + BALCHQ_IND + BALSAV_IND +
+                 BALLOAN_IND + BALLOC_IND + BALMRGG_IND,
+                 data = filter(rrsp2, Sample == "Estimation"),
+                 family = binomial(logit))
> # Print summary
> summary(rrsplgis) # AIC = 1586.6

Call:
glm(formula = APURCH ~ age + gendm + atmcrd + paydep + BALCHQ +
    BALSAV + TOTDEP + BALLOAN + BALLOC + BALMRGG + NEWLOC + NEWMRGG +
    TXBRAN + TXATM + TXPOS + TXCHQ + TXWEB + TXTEL + TOTSERV.NEW +
    valsegm + numrr_1 + numcon_1 + avginc_1 + avginv_1 + BALCHQ_IND +
    BALSAV_IND + BALLOAN_IND + BALLOC_IND + BALMRGG_IND, family = binomial(logit),
    data = filter(rrsp2, Sample == "Estimation"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.7888  -0.4923   0.0130   0.3613   3.4852

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.418e+01  1.101e+00  21.956 < 2e-16 ***
age          -6.232e-03  6.290e-03  -0.991 0.321816
gendm1       -2.192e-01  1.377e-01  -1.592 0.111483
atmcrd1       2.397e-01  1.882e-01   1.274 0.202763
paydep1       5.579e-01  1.604e-01   3.478 0.000505 ***
BALCHQ       -2.403e-05  1.546e-05  -1.554 0.120141
BALSAV       1.589e-05  2.832e-05   0.561 0.574865
TOTDEP       5.185e-05  4.738e-06  10.943 < 2e-16 ***
BALLOAN      2.649e-05  1.972e-05   1.344 0.179011
BALLOC      -1.791e-06  3.170e-06  -0.565 0.571967
BALMRGG     -1.833e-07  1.800e-06  -0.102 0.918906
NEWLOC1     -2.055e+00  4.262e-01  -4.822 1.42e-06 ***
NEWMRGG1    -6.097e-01  5.586e-01  -1.092 0.275004
TXBRAN       1.033e-01  2.697e-02   3.829 0.000128 ***
TXATM        1.079e-02  1.702e-02   0.634 0.526393
TXPOS       -5.220e-03  6.440e-03  -0.811 0.417619
TXCHQ       -1.080e-02  2.164e-02  -0.499 0.617722
TXWEB       -1.172e-02  3.126e-02  -0.375 0.707668
TXTEL       -1.955e+00  1.506e+00  -1.298 0.194227
TOTSERV.NEW -4.787e+00  1.838e-01  -26.041 < 2e-16 ***
valsegmA     1.057e+00  4.679e-01   2.259 0.023875 *
valsegmB     1.178e+00  4.412e-01   2.671 0.007561 **
valsegmC     1.327e+00  3.432e-01   3.865 0.000111 ***
valsegmD     1.634e+00  2.229e-01   7.331 2.29e-13 ***
numrr_1     -9.934e-04  4.830e-04  -2.057 0.039730 *
numcon_1     2.161e-04  1.042e-04   2.074 0.038098 *
avginc_1    -1.174e-06  9.475e-06  -0.124 0.901405
avginv_1     1.109e-05  2.730e-05   0.406 0.684579
BALCHQ_IND1 -4.443e+00  2.957e-01  -15.025 < 2e-16 ***
BALSAV_IND1 -4.861e+00  2.425e-01  -20.042 < 2e-16 ***
BALLOAN_IND1 -3.802e+00  2.767e-01  -13.741 < 2e-16 ***
BALLOC_IND1 -4.592e+00  2.396e-01  -19.171 < 2e-16 ***
BALMRGG_IND1 -4.568e+00  3.874e-01  -11.792 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3450.1  on 2488  degrees of freedom
Residual deviance: 1520.6  on 2456  degrees of freedom
AIC: 1586.6

Number of Fisher Scoring iterations: 6

```

Figure 8: Results of the First Logistic Model

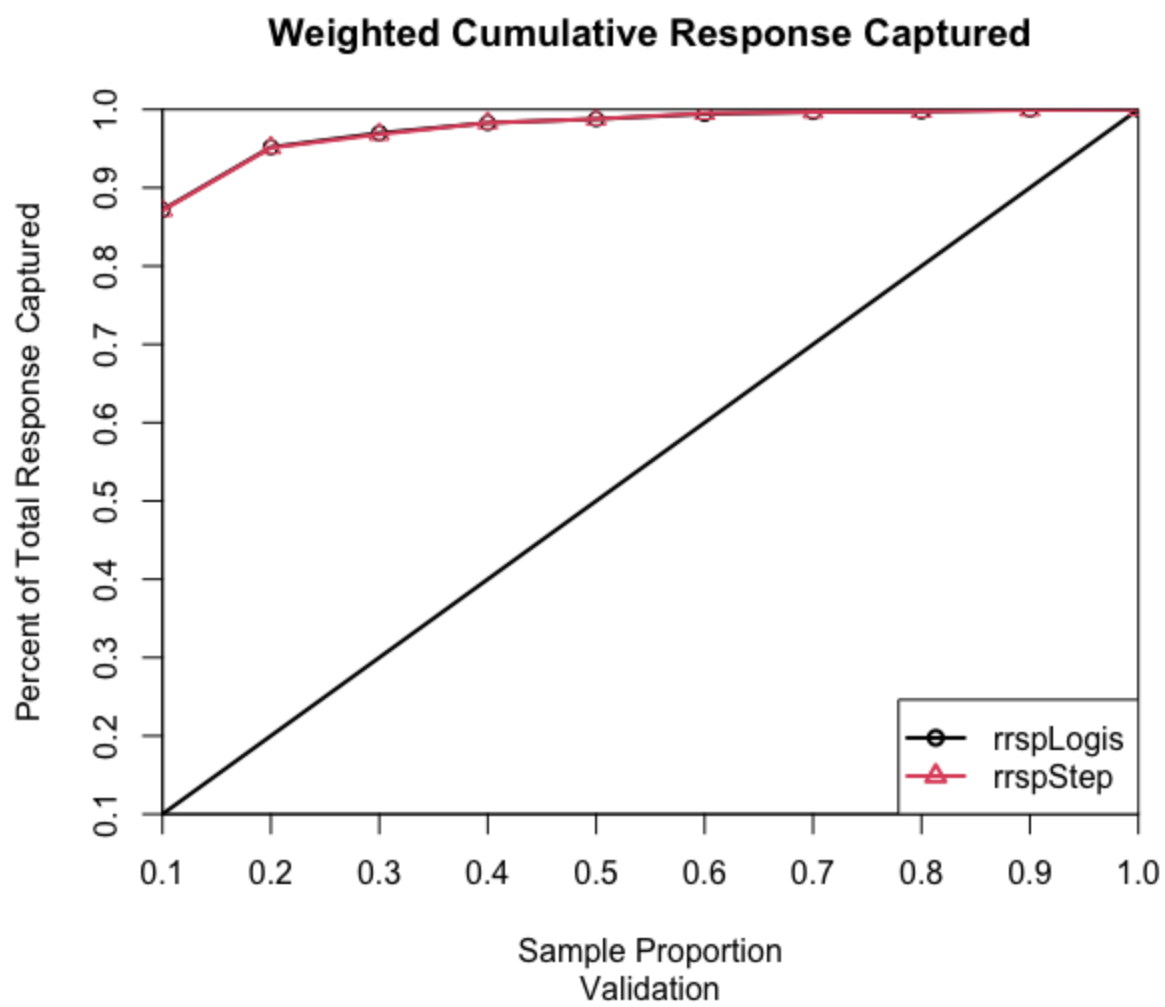


Figure 9: The Lift Chart of the First Logistic Model


```

520 # Now we can remove both TOTSERV and TOTSERV.NEW from our subsequent modelling
521 # Rework and overwrite the previous models
522 rrspForestAll <- randomForest(formula = APURCH ~ age + gendm + atmcrd + paydep +
523                               BALCHQ + BALSAV + TOTDEP + BALLOAN + BALLOC +
524                               BALMRGG + NEWLOC + NEWMRGG + TXBRAN + TXATM +
525                               TXPOS + TXCHQ + TXWEB + TXTEL + valsegm +
526                               numrr_1 + numcon_1 + avginc_1 + avginv_1 +
527                               BALCHQ_IND + BALSAV_IND + BALLOAN_IND +
528                               BALLOC_IND + BALMRGG_IND,
529                               data = filter(rrsp2, Sample == "Estimation"),
530                               importance = TRUE,
531                               ntree = 500, mtry = 4)
532
533 # Contingency Table
534 rrspForestAll[["confusion"]]
535 # Misclassification rate from the internal "out of bag" validation set is 0.39
536 # for "N" and 0.35 for "Y". The overall hit rate is 63%, or equivalently,
537 # the error rate is 37%. That's only a 13% improvement over the 50% hit rate we
538 # would get without a model or the 50% response rate we would get
539 # if we had to randomly target individuals in a 50/50 database.
540
541 # Plot the relative importance of the variables:
542 varImpPlot(rrspForestAll, type = 2,
543            main = "rrspForestAll", # title
544            cex = 0.7) # font size
545 # Now the importance plot shows TOTDEP as the most predictive feature.
546
545:71 # Logistic Regression

```

Console Terminal Background Jobs

R 4.1.2 · ~/Desktop/sfu/s13/BUS445/Assignment2/ ↗

```

> # Now we can remove both TOTSERV and TOTSERV.NEW from our subsequent modelling
> # Rework and overwrite the previous models
> rrspForestAll <- randomForest(formula = APURCH ~ age + gendm + atmcrd + paydep +
+                               BALCHQ + BALSAV + TOTDEP + BALLOAN + BALLOC +
+                               BALMRGG + NEWLOC + NEWMRGG + TXBRAN + TXATM +
+                               TXPOS + TXCHQ + TXWEB + TXTEL + valsegm +
+                               numrr_1 + numcon_1 + avginc_1 + avginv_1 +
+                               BALCHQ_IND + BALSAV_IND + BALLOAN_IND +
+                               BALLOC_IND + BALMRGG_IND,
+                               data = filter(rrsp2, Sample == "Estimation"),
+                               importance = TRUE,
+                               ntree = 500, mtry = 4)
> # Contingency Table
> rrspForestAll[["confusion"]]
  N   Y class.error
N 749 480   0.3905614
Y 441 819   0.3500000

```

Figure 10: The New Random Forest without TOTSERV

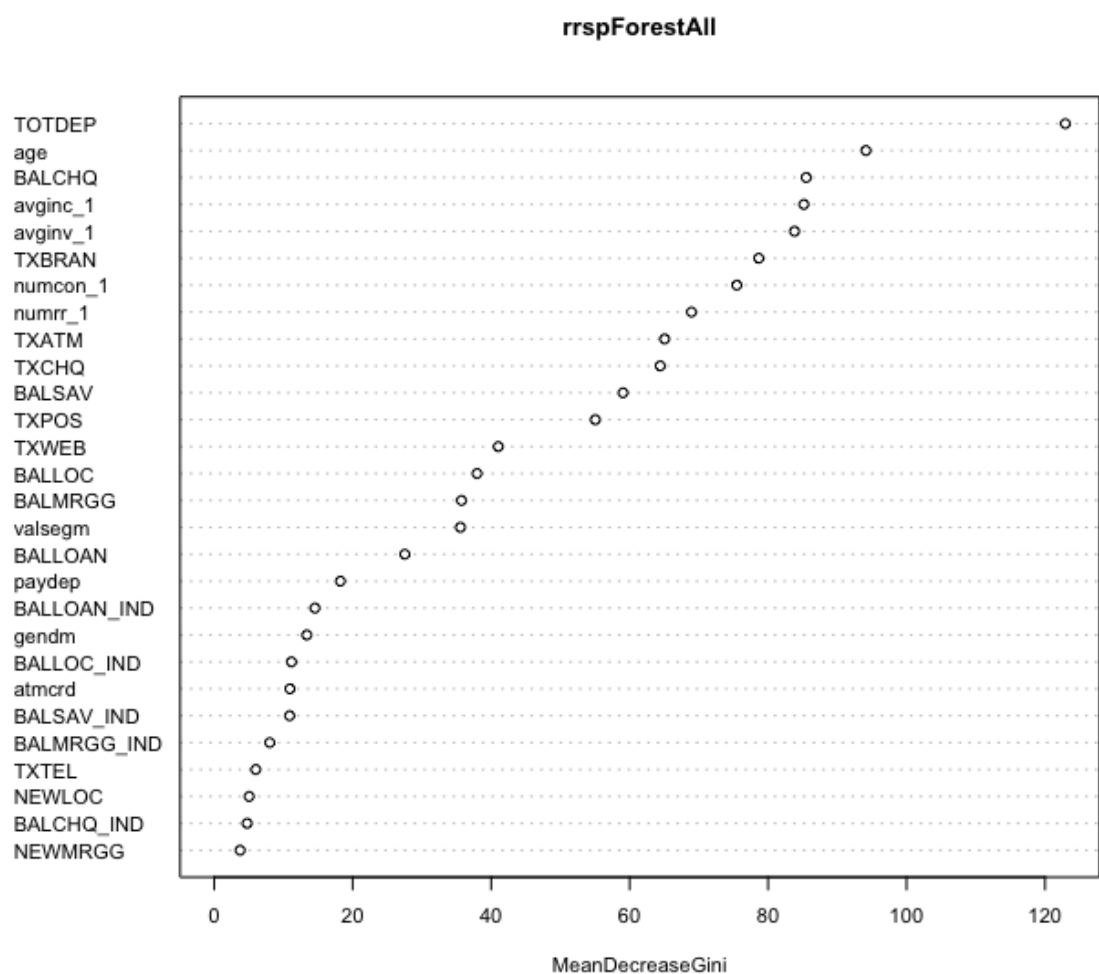


Figure 11: The New Importance Plot without TOTSERV

```

> # Create a logistic regression model
> rrsplLogis <- glm(formula = APURCH ~ age + gendm + atmcrd + paydep + BALCHQ +
+                   BALSAV + TOTDEP + BALLOAN + BALLOC + BALMRGG + NEWLOC +
+                   NEWMRGG + TXBRAN + TXATM + TXPOS + TXCHQ + TXWEB + TXTEL +
+                   valsegm + numrr_1 + numcon_1 + avginc_1 + avginv_1 +
+                   BALCHQ_IND + BALSAV_IND + BALLOAN_IND + BALLOC_IND + BALMRGG_IND,
+                   data = filter(rrsp2, Sample == "Estimation"),
+                   family = binomial(logit))
> # Print summary
> summary(rrsplLogis) # AIC = 3260.2

Call:
glm(formula = APURCH ~ age + gendm + atmcrd + paydep + BALCHQ +
    BALSAV + TOTDEP + BALLOAN + BALLOC + BALMRGG + NEWLOC + NEWMRGG +
    TXBRAN + TXATM + TXPOS + TXCHQ + TXWEB + TXTEL + valsegm +
    numrr_1 + numcon_1 + avginc_1 + avginv_1 + BALCHQ_IND + BALSAV_IND +
    BALLOAN_IND + BALLOC_IND + BALMRGG_IND, family = binomial(logit),
    data = filter(rrsp2, Sample == "Estimation"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1048  -1.0726   0.5249   1.0912   1.8658

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.336e-01  3.969e-01   0.841 0.400583
age          -1.937e-02  3.989e-03  -4.856 1.20e-06 ***
gendm1       2.693e-02  8.750e-02   0.308 0.758217
atmcrd1      3.061e-01  1.197e-01   2.557 0.010561 *
paydep1      4.733e-01  1.031e-01   4.590 4.43e-06 ***
BALCHQ       1.835e-05  9.151e-06   2.005 0.044920 *
BALSAV       6.799e-05  2.157e-05   3.152 0.001624 **
TOTDEP       5.749e-06  2.245e-06   2.561 0.010451 *
BALLOAN      9.091e-06  1.244e-05   0.731 0.465014
BALLOC       1.905e-06  1.927e-06   0.989 0.322849
BALMRGG      2.483e-06  1.145e-06   2.168 0.030189 *
NEWLOC1      4.233e-01  2.726e-01   1.553 0.120451
NEWMRGG1     2.806e-01  3.428e-01   0.819 0.413016
TXBRAN       2.939e-02  1.621e-02   1.813 0.069820 .
TXATM        4.476e-03  1.100e-02   0.407 0.684026
TXPOS       -2.678e-03  4.117e-03  -0.650 0.515372
TXCHQ        5.111e-03  1.510e-02   0.338 0.735023
TXWEB        1.543e-02  2.006e-02   0.769 0.441729
TXTEL       -1.224e+00  9.174e-01  -1.334 0.182217
valsegmA     1.010e-02  2.827e-01   0.036 0.971501
valsegmB    -6.974e-02  2.629e-01  -0.265 0.790838
valsegmC    -4.715e-02  2.076e-01  -0.227 0.820322
valsegmD     4.263e-01  1.298e-01   3.283 0.001026 **
numrr_1     -1.904e-04  3.209e-04  -0.593 0.552924
numcon_1     3.608e-05  6.963e-05   0.518 0.604411
avginc_1     8.314e-06  6.214e-06   1.338 0.180950
avginv_1    -1.121e-05  1.876e-05  -0.597 0.550235
BALCHQ_IND1  1.035e-01  1.535e-01   0.674 0.500048
BALSAV_IND1  4.349e-02  9.548e-02   0.455 0.648783
BALLOAN_IND1 -5.959e-01  1.605e-01  -3.713 0.000205 ***
BALLOC_IND1 -1.199e-01  1.041e-01  -1.151 0.249739
BALMRGG_IND1 -1.817e-01  2.132e-01  -0.852 0.394124
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3450.1  on 2488  degrees of freedom
Residual deviance: 3196.2  on 2457  degrees of freedom
AIC: 3260.2

Number of Fisher Scoring iterations: 4

```

Figure 12: Results of the New Maximal Logistic Model

```

563 # Run a stepwise regression using the "rrsplogis" model
564 rrspStep <- step(rrspLogis, direction = "both")
565 summary(rrspStep) # AIC = 3238.8
566
567:1 # Logistic Regression

```

Console Terminal Background Jobs

R 4.1.2 · ~/Desktop/sfu/s13/BUS445/Assignment2/ ↗

```

> summary(rrspStep) # AIC = 3238.8

Call:
glm(formula = APURCH ~ age + atmcrd + paydep + BALCHQ + BALSAV +
    TOTDEP + BALMRGG + NEWLOC + TXBRAN + valsegm + avginc_1 +
    BALLOAN_IND + BALLOC_IND, family = binomial(logit), data = filter(rrsp2,
    Sample == "Estimation"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.028  -1.075   0.545   1.093   1.847

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.001e-01  2.950e-01   1.017  0.308946
age          -1.894e-02  3.885e-03  -4.875  1.09e-06 ***
atmcrd1       3.047e-01  1.130e-01   2.695  0.007036 **
paydep1       4.770e-01  9.637e-02   4.949  7.45e-07 ***
BALCHQ        1.867e-05  9.060e-06   2.061  0.039307 *
BALSAV        6.820e-05  2.070e-05   3.295  0.000983 ***
TOTDEP        5.400e-06  2.170e-06   2.489  0.012822 *
BALMRGG       3.260e-06  8.128e-07   4.011  6.05e-05 ***
NEWLOC1       4.932e-01  2.659e-01   1.855  0.063624 .
TXBRAN        2.480e-02  1.529e-02   1.622  0.104767
valsegmA      1.115e-01  2.388e-01   0.467  0.640461
valsegmB     -1.032e-03  2.309e-01  -0.004  0.996435
valsegmC      7.855e-03  1.865e-01   0.042  0.966406
valsegmD      4.236e-01  1.279e-01   3.313  0.000924 ***
avginc_1      6.439e-06  4.206e-06   1.531  0.125747
BALLOAN_IND1 -6.621e-01  1.192e-01  -5.555  2.78e-08 ***
BALLOC_IND1  -1.414e-01  9.960e-02  -1.419  0.155828
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3450.1  on 2488  degrees of freedom
Residual deviance: 3204.8  on 2472  degrees of freedom
AIC: 3238.8

Number of Fisher Scoring iterations: 4

```

Figure 13: Results of the New Stepwise Regression Model

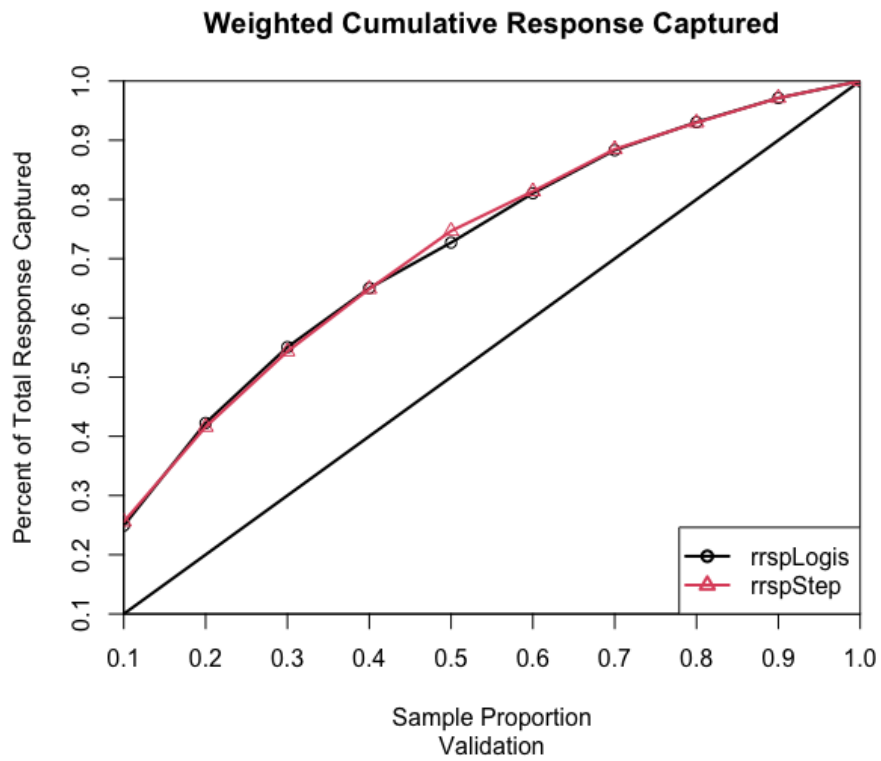


Figure 14: Compare the Maximal and Stepwise Model Performance

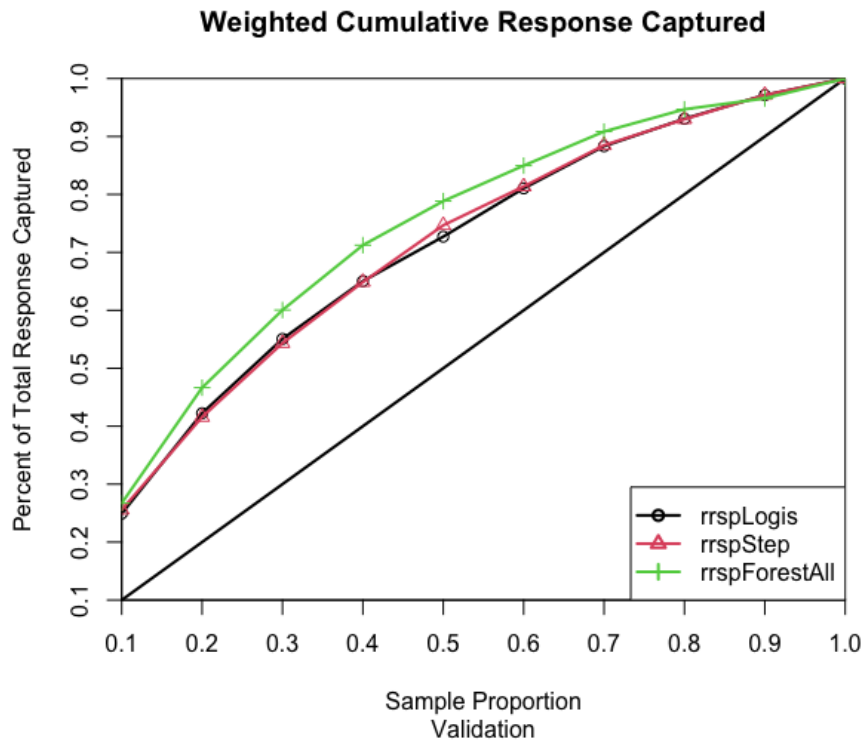


Figure 15: Compare the Logistic Models with the Random Forest

<i>Rank</i>	<i>rrspForestAllv</i>	<i>rrspStep (by significance)</i>
1	TOTDEP	BALLOAN.IND1
2	age	paydep1
3	BALCHQ	age
4	avginc_1	BALMRGG
5	avginv_1	valsegmD
6	TXBRAN	BALSAV
7	numcon_1	atmcrd1
8	numrr_1	TOTDEP
9	TXATM	BALCHQ
10	TXCHQ	NEWLOC1
11	BALSAV	TXBRAN
12	TXPOS	avcinc_1
13	TXWEB	BALLOC_IND1
14	BALLOC	valsegmA
15	BALMRGG	valsegmC

Figure 16: Compare Important Variables of Forest and Stepwise Model

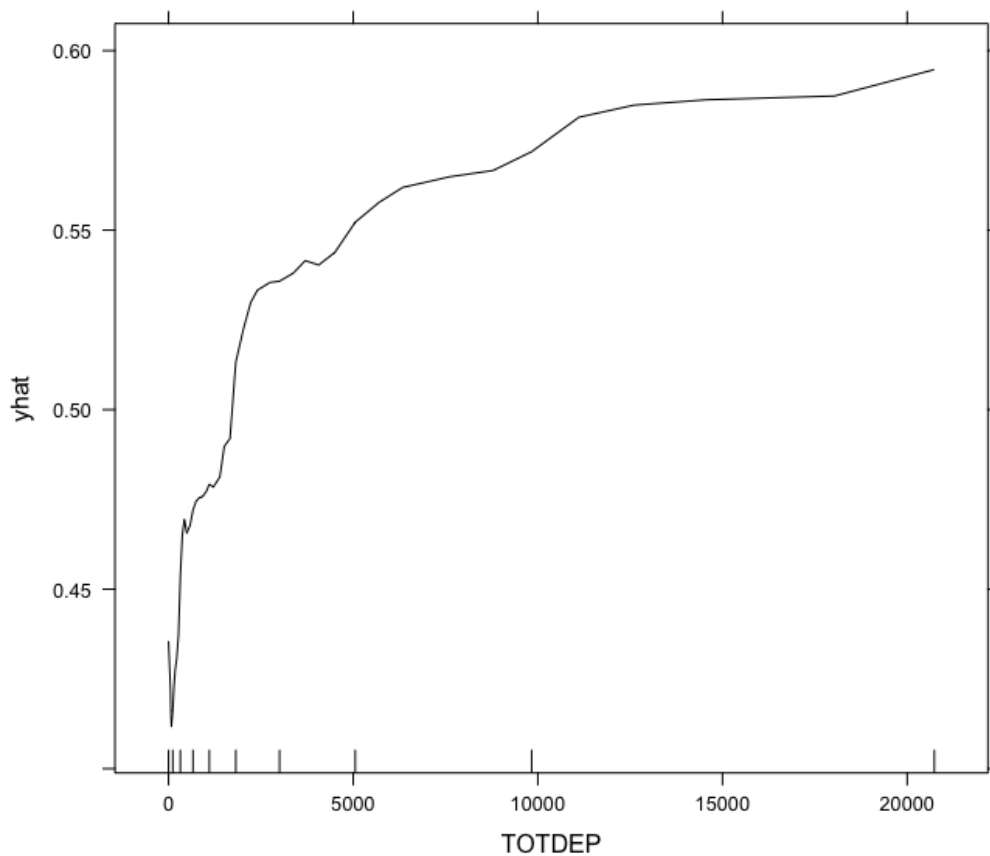



Figure 17: The Partial Dependence Plot of *TOTDEP*

```

841 ##### Make a table of comparison #####
842
843 p_comparison <- data.frame(Linear = c(as.data.frame(coef(summary(rrspLogis))[,4])[-1])[,1][1:26],NA,NA,
844                                     as.data.frame(coef(summary(rrspLogis))[,4])[-1])[,1][27:31]),
845                               Mixed1 = c(as.data.frame(coef(summary(rrspMixed1))[,4])[-1])[,1][1:22],NA,NA,NA,NA,
846                                           as.data.frame(coef(summary(rrspMixed1))[,4])[-1])[,1][23:29]),
847                               Mixed2 = c(as.data.frame(coef(summary(rrspMixed2))[,4])[-1])[,1][1:22],NA,NA,NA,NA,
848                                           as.data.frame(coef(summary(rrspMixed2))[,4])[-1])[,1][23:29]),
849                               Mixed3 = c(as.data.frame(coef(summary(rrspMixed3))[,4])[-1])[,1][1:22],NA,NA,NA,NA,
850                                           as.data.frame(coef(summary(rrspMixed3))[,4])[-1])[,1][23:29]),
851                               row.names = c(row.names(as.data.frame(coef(summary(rrspLogis))[,4])[-1])[1:26],
852                                              'PC.numcon.numrr', 'PC.avginc.avginv',
853                                              row.names(as.data.frame(coef(summary(rrspLogis))[,4])[-1])[27:31]))
854 # To check the effect of Logged TOTDEP and 2 PCA's, compare column 1 and 2;
855 # To check the effect of Logged BAL, compare column 2 and 3;
856 # To check the effect of Logged TX, compare column 3 and 4;
857 p_comparison

```

859:1  Make a table of comparison :

Console Terminal × Background Jobs ×

R 4.1.2 · ~/Desktop/sfu/s13/BUS445/Assignment2/ ➔

	Linear	Mixed1	Mixed2	Mixed3
age	1.196343e-06	7.412621e-09	1.515868e-08	8.234901e-09
gendm1	7.582166e-01	7.164540e-01	6.532656e-01	6.864725e-01
atmcrd1	1.056071e-02	2.162748e-02	2.644823e-02	3.645243e-02
paydep1	4.429735e-06	3.976989e-04	3.414805e-04	5.062869e-04
BALCHQ	4.492010e-02	6.905415e-01	6.656227e-01	4.854646e-01
BALSAV	1.623980e-03	1.604410e-01	9.825321e-02	1.018776e-01
TOTDEP	1.045053e-02	2.566350e-17	2.545779e-14	7.850367e-14
BALLOAN	4.650142e-01	4.920194e-01	8.970931e-01	8.942973e-01
BALLOC	3.228494e-01	3.773442e-01	6.027712e-02	9.662740e-02
BALMRGG	3.018928e-02	5.427123e-02	6.848024e-02	6.431234e-02
NEWLOC1	1.204510e-01	2.585924e-01	2.549872e-01	2.512405e-01
NEWMRGG1	4.130159e-01	6.613914e-01	6.983680e-01	7.038711e-01
TXBRAN	6.981954e-02	1.411367e-01	1.495835e-01	4.417158e-02
TXATM	6.840261e-01	6.172741e-01	6.392165e-01	7.598052e-01
TXPOS	5.153722e-01	6.843255e-01	6.422597e-01	3.578050e-01
TXCHQ	7.350234e-01	9.204901e-01	9.276362e-01	4.414443e-01
TXWEB	4.417293e-01	5.021056e-01	5.963510e-01	3.669939e-01
TXTEL	1.822172e-01	1.914839e-01	1.727692e-01	1.625999e-01
valsegmA	9.715011e-01	4.142350e-01	3.037139e-01	2.616823e-01
valsegmB	7.908377e-01	1.350898e-01	8.316219e-02	7.329936e-02
valsegmC	8.203222e-01	7.475950e-02	3.791421e-02	3.077959e-02
valsegmD	1.026121e-03	6.958528e-01	8.980363e-01	9.593043e-01
numrr_1	5.529245e-01	NA	NA	NA
numcon_1	6.044108e-01	NA	NA	NA
avginc_1	1.809501e-01	NA	NA	NA
avginv_1	5.502351e-01	NA	NA	NA
PC.numcon.numrr	NA	3.382895e-01	3.269021e-01	3.453956e-01
PC.avginc.avginv	NA	3.122647e-01	3.143678e-01	3.508963e-01
BALCHQ_IND1	5.000480e-01	2.550377e-01	6.915690e-01	7.157868e-01
BALSAV_IND1	6.487825e-01	2.266537e-01	6.024181e-02	5.609847e-02
BALLOAN_IND1	2.045473e-04	6.795537e-05	3.273696e-01	3.234848e-01
BALLOC_IND1	2.497394e-01	1.200025e-02	5.982481e-01	6.055940e-01
BALMRGG_IND1	3.941238e-01	3.708793e-01	1.300173e-01	1.225177e-01

Figure 18: Compare the Significance Levels of Variables


```
> # Build the model based on the variables we identified above:
> rrspMixed4 <- glm(formula = APURCH ~ age + gendm + atmcrd + paydep + Log.BALCHQ +
+ Log.BALSAV + Log.TOTDEP + BALLOAN + Log.BALLOC + BALMRGG +
+ NEWLOC + NEWMRGG + Log.TXBRAN + TXATM + Log.TXPOS + Log.TXCHQ +
+ Log.TXWEB + Log.TXTEL + valsegm +
+ PC.numcon.numrr + PC.avginc.avginv + BALCHQ_IND + BALSAB_IND +
+ BALLOAN_IND + BALLOC_IND + BALMRGG_IND,
+ data = filter(rrsp2, Sample == "Estimation"),
+ family = binomial(logit))
> summary(rrspMixed4) # AIC = 3182.5 (better than Mixed3)
```

Call:

```
glm(formula = APURCH ~ age + gendm + atmcrd + paydep + Log.BALCHQ +
Log.BALSAV + Log.TOTDEP + BALLOAN + Log.BALLOC + BALMRGG +
NEWLOC + NEWMRGG + Log.TXBRAN + TXATM + Log.TXPOS + Log.TXCHQ +
Log.TXWEB + Log.TXTEL + valsegm + PC.numcon.numrr + PC.avginc.avginv +
BALCHQ_IND + BALSAB_IND + BALLOAN_IND + BALLOC_IND + BALMRGG_IND,
family = binomial(logit), data = filter(rrsp2, Sample ==
"Estimation"))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.2659	-1.0610	0.4931	1.0539	2.1709

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-8.245e-01	3.994e-01	-2.064	0.038976 *
age	-2.407e-02	4.138e-03	-5.816	6.03e-09 ***
gendm1	3.118e-02	8.893e-02	0.351	0.725907
atmcrd1	2.868e-01	1.255e-01	2.285	0.022334 *
paydep1	3.809e-01	1.071e-01	3.557	0.000375 ***
Log.BALCHQ	-1.762e-02	2.496e-02	-0.706	0.480294
Log.BALSAV	3.241e-02	1.966e-02	1.648	0.099260 .
Log.TOTDEP	2.644e-01	3.539e-02	7.472	7.91e-14 ***
BALLOAN	9.444e-06	1.306e-05	0.723	0.469792
Log.BALLOC	3.468e-02	2.095e-02	1.655	0.097883 .
BALMRGG	2.192e-06	1.174e-06	1.867	0.061856 .
NEWLOC1	3.128e-01	2.765e-01	1.132	0.257806
NEWMRGG1	1.233e-01	3.443e-01	0.358	0.720166
Log.TXBRAN	1.364e-01	6.925e-02	1.970	0.048842 *
TXATM	5.684e-03	1.140e-02	0.498	0.618146
Log.TXPOS	-4.854e-02	4.857e-02	-0.999	0.317667
Log.TXCHQ	5.371e-02	7.452e-02	0.721	0.471030
Log.TXWEB	6.449e-02	7.252e-02	0.889	0.373832
Log.TXTEL	-1.678e+00	1.175e+00	-1.428	0.153326
valsegmA	-3.508e-01	2.820e-01	-1.244	0.213568
valsegmB	-4.721e-01	2.687e-01	-1.757	0.078976 .
valsegmC	-4.626e-01	2.140e-01	-2.162	0.030640 *
valsegmD	1.063e-02	1.401e-01	0.076	0.939530
PC.numcon.numrr	-2.792e-02	2.894e-02	-0.964	0.334800
PC.avginc.avginv	-3.027e-02	3.340e-02	-0.907	0.364662
BALCHQ_IND1	7.352e-02	2.027e-01	0.363	0.716800
BALSAV_IND1	2.318e-01	1.205e-01	1.924	0.054391 .
BALLOAN_IND1	-6.471e-01	1.656e-01	-3.908	9.33e-05 ***
BALLOC_IND1	-7.951e-02	1.480e-01	-0.537	0.591050
BALMRGG_IND1	-2.200e-01	2.150e-01	-1.023	0.306101

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3450.1 on 2488 degrees of freedom
Residual deviance: 3122.5 on 2459 degrees of freedom
AIC: 3182.5

Number of Fisher Scoring iterations: 4

Figure 19: Mixed4 Maximal Regression Model


```

> summary(rrspStep.Mixed4) # AIC = 3163.5 (Better than Mixed3)

Call:
glm(formula = APURCH ~ age + atmcrd + paydep + Log.BALSAV + Log.TOTDEP +
     Log.BALLOC + BALMRGG + Log.TXBRAN + Log.TXTEL + valsegm +
     BALSAV_IND + BALLOAN_IND, family = binomial(logit), data = filter(rrsp2,
     Sample == "Estimation"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3739  -1.0572   0.4989   1.0566   2.1317

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.016e+00  2.743e-01  -3.703 0.000213 ***
age          -2.309e-02  3.969e-03  -5.818 5.96e-09 ***
atmcrd1       2.648e-01  1.144e-01   2.314 0.020695 *
paydep1       3.849e-01  9.715e-02   3.962 7.44e-05 ***
Log.BALSAV    3.768e-02  1.855e-02   2.031 0.042208 *
Log.TOTDEP    2.544e-01  2.798e-02   9.093 < 2e-16 ***
Log.BALLOC    5.025e-02  1.369e-02   3.669 0.000243 ***
BALMRGG       3.137e-06  8.232e-07   3.810 0.000139 ***
Log.TXBRAN    1.266e-01  6.512e-02   1.944 0.051848 .
Log.TXTEL     -1.758e+00  1.169e+00  -1.503 0.132743
valsegmA     -3.263e-01  2.529e-01  -1.290 0.196943
valsegmB     -4.017e-01  2.385e-01  -1.684 0.092137 .
valsegmC     -4.050e-01  1.934e-01  -2.094 0.036267 *
valsegmD       6.203e-03  1.369e-01   0.045 0.963864
BALSAV_IND1   2.240e-01  1.193e-01   1.879 0.060290 .
BALLOAN_IND1 -7.320e-01  1.215e-01  -6.025 1.69e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3450.1  on 2488  degrees of freedom
Residual deviance: 3131.5  on 2473  degrees of freedom
AIC: 3163.5

Number of Fisher Scoring iterations: 4

```

Figure 20: Mixed4 Stepwise Regression Model

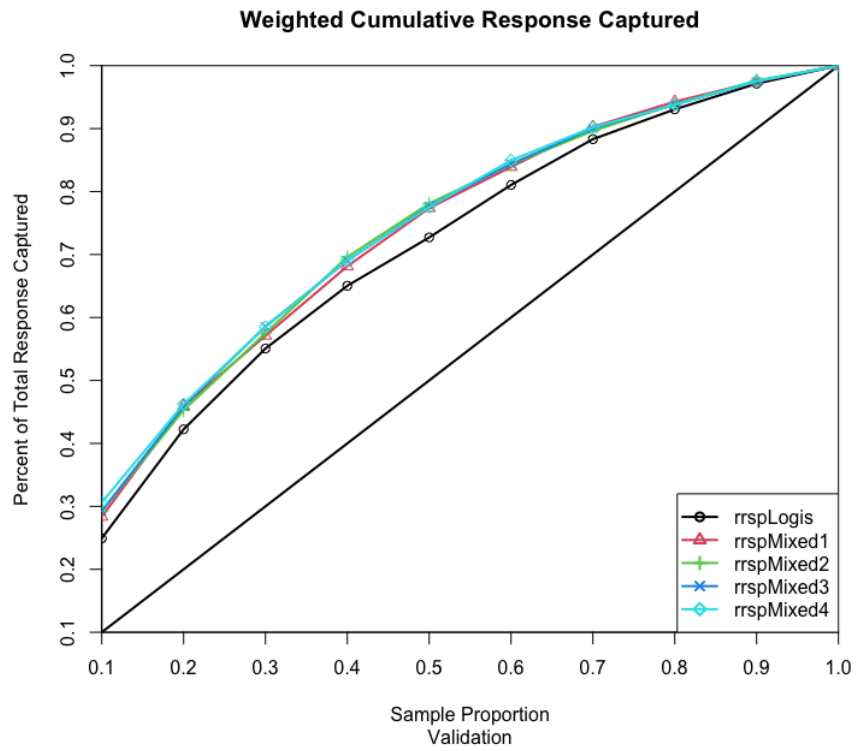


Figure 21: Compare Four Mixed Maximal Models with the First Logistic Maximal Model

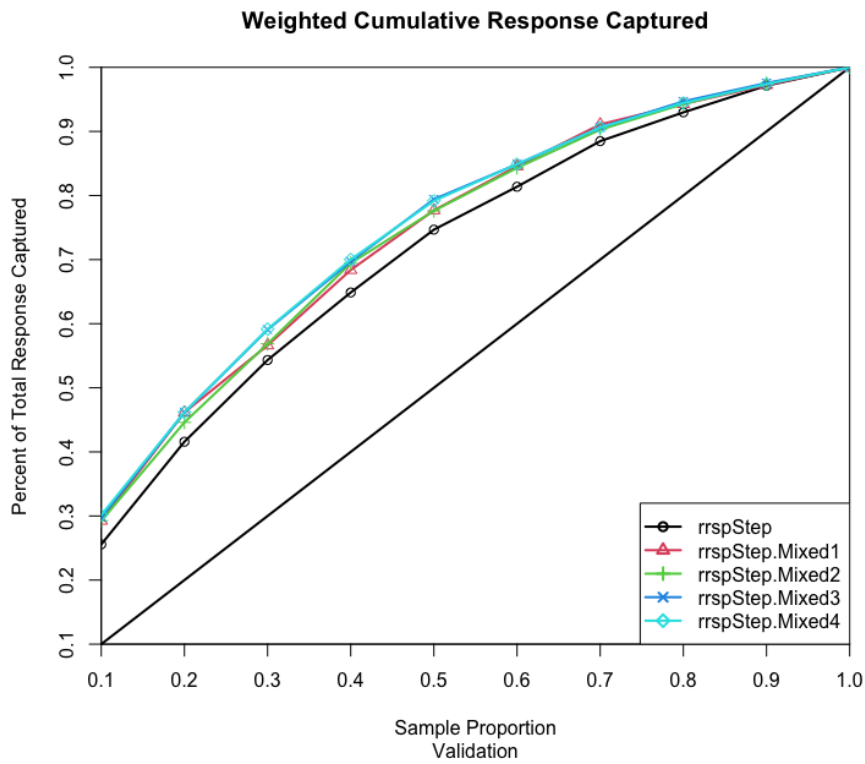


Figure 22: Compare Four Mixed Stepwise Models with the First Stepwise Model

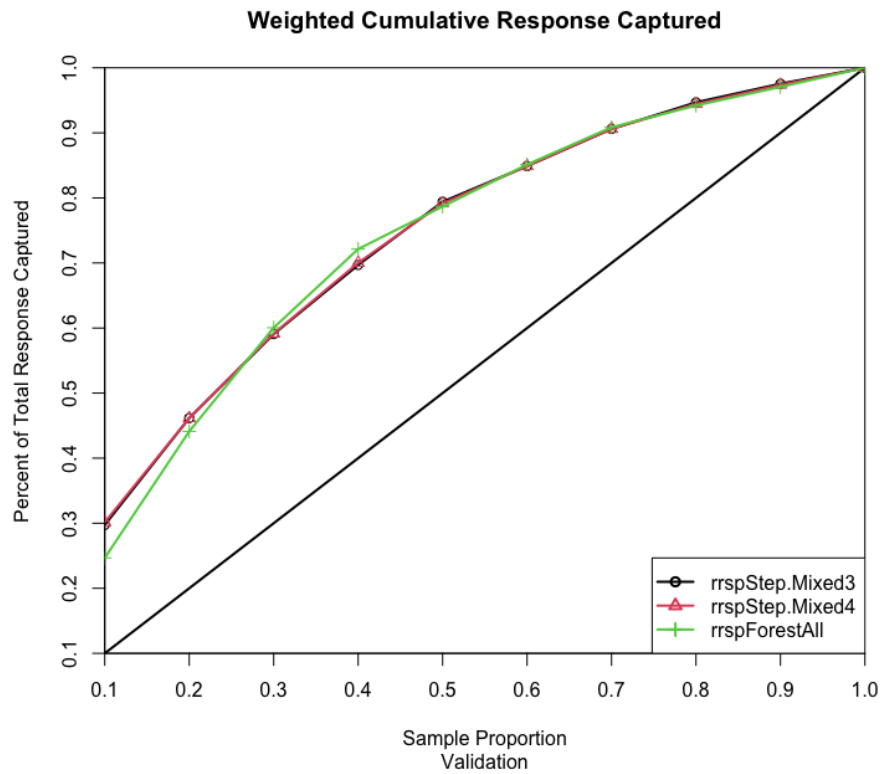


Figure 23: Compare the Best Two Stepwise Models with the Random Forest

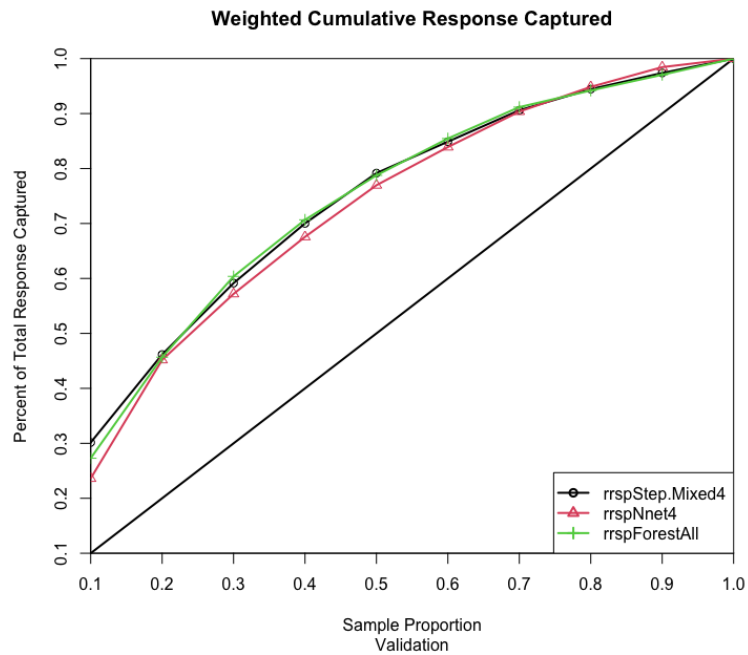


Figure 24: Compare the Best Stepwise Model, Random Forest, and Neural Network

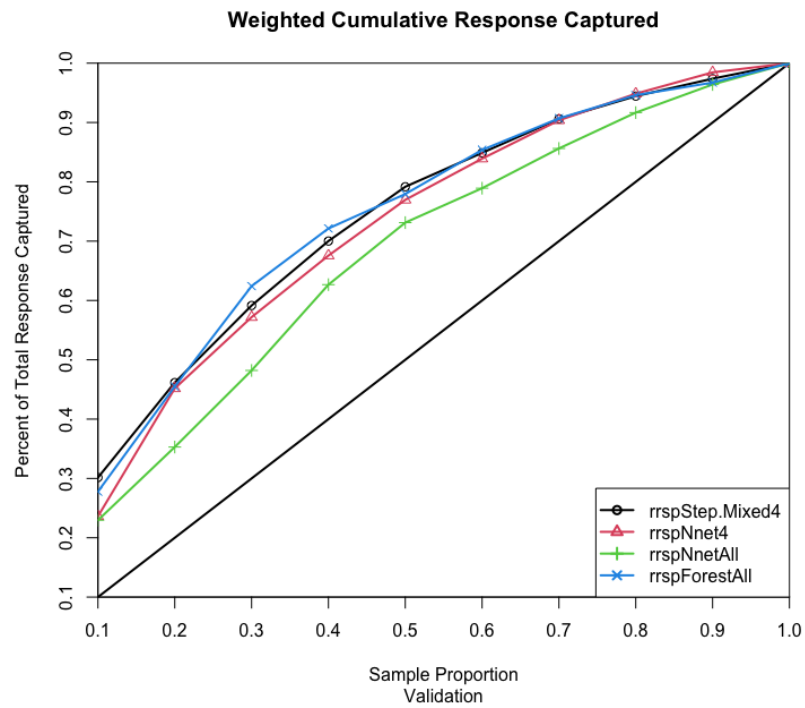
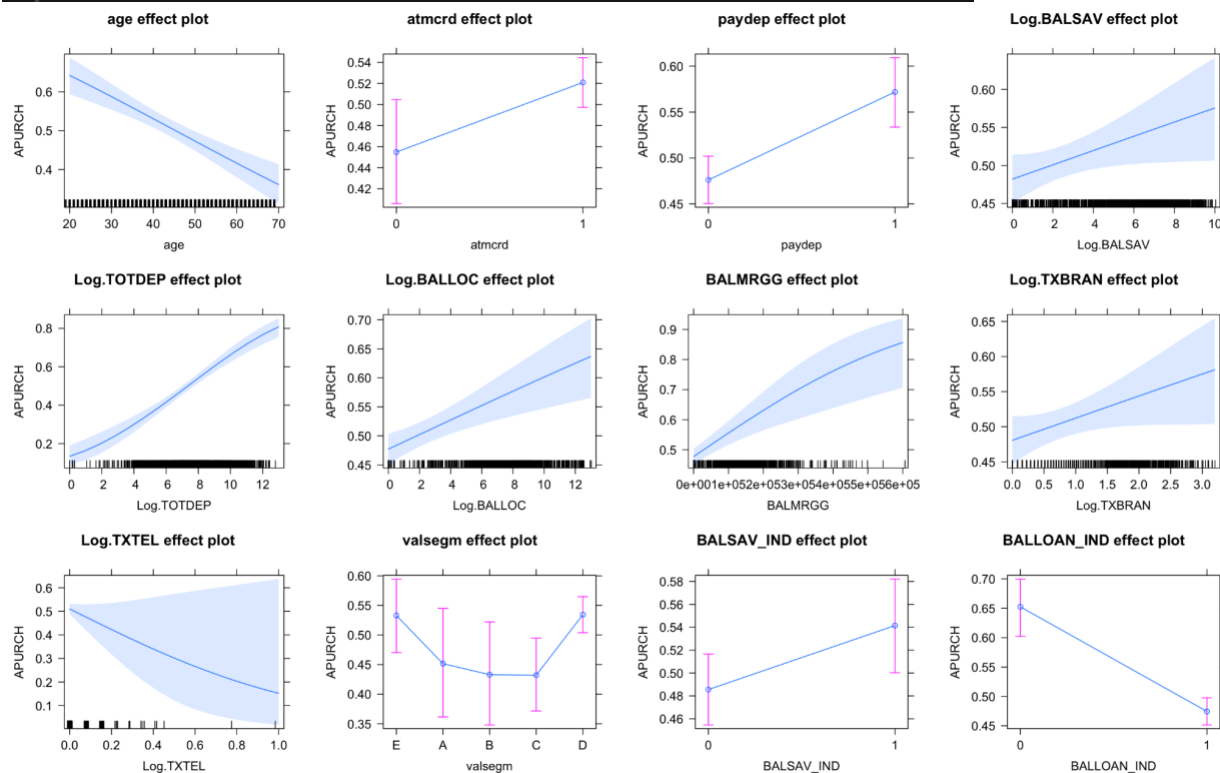


Figure 25: Compare the Best Stepwise Model, Random Forest, Neural Network and Neural Network All

Question 2

	Variable	Coefficient	StdError	Zval	Pval
1	Log.TOTDEP	2.544011e-01	2.797678e-02	9.09329264	9.608420e-20
2	BALLOAN_IND1	-7.319873e-01	1.214861e-01	-6.02527443	1.688229e-09
3	age	-2.309337e-02	3.969293e-03	-5.81800550	5.955394e-09
4	paydep	3.848939e-01	9.714941e-02	3.96187547	7.436333e-05
5	BALMRGG	3.136563e-06	8.232191e-07	3.81011894	1.389000e-04
6	Log.BALLOC	5.024757e-02	1.369459e-02	3.66915396	2.433545e-04
7	atmcdr1	2.647710e-01	1.144461e-01	2.31350029	2.069515e-02
8	valsegmC	-4.049877e-01	1.934112e-01	-2.09392071	3.626704e-02
9	Log.BALSAV	3.768271e-02	1.854958e-02	2.03145902	4.220845e-02
10	Log.TXBRAN	1.266208e-01	6.512093e-02	1.94439416	5.184794e-02
11	BALSAV_IND1	2.240476e-01	1.192587e-01	1.87866795	6.028985e-02
12	valsegmB	-4.017319e-01	2.385252e-01	-1.68423295	9.213666e-02
13	Log.TXTEL	-1.758072e+00	1.169420e+00	-1.50337149	1.327433e-01
14	valsegmA	-3.262942e-01	2.528805e-01	-1.29030960	1.969432e-01
15	valsegmD	6.202797e-03	1.369105e-01	0.04530547	9.638638e-01

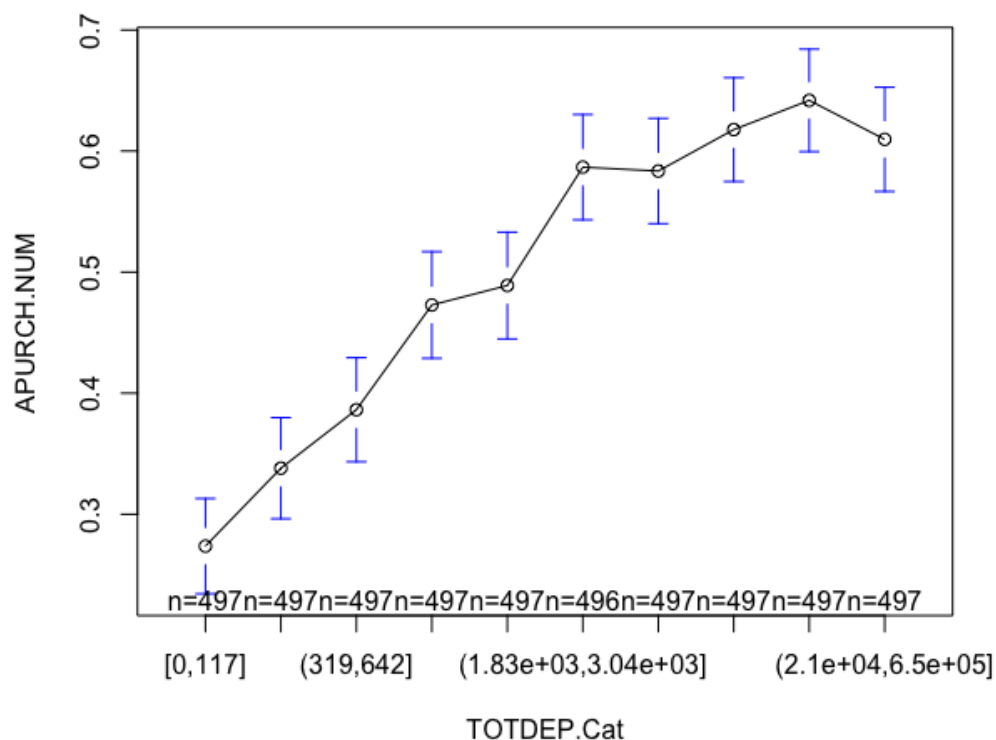


In our mental model, we believe that three factors are playing important roles in determining whether an existing client would be an RRSP purchaser. They are (1) the capability of purchases and the needs of an RRSP account, (2) the relationship with VanCity and customer intimacy, and (3) the frequency interacting with VanCity where a higher frequency is likely to lead to higher exposure to VanCity campaigns, products, and investment discussions. Based on our final model, we selected 5 most important predictors considering their significance levels (p-values),

the sizes of their effects (coefficients), and whether they are intuitive and consistent with our mental model.

1. *TOTDEP*: the average total monthly deposits over previous 12 months

This feature is believed to be an important predictor because of its highest significance, relatively high magnitude of impact, and a large effect size with a high confidence level. For every 1% increase in the total deposit, we expect the purchase possibility to increase by 0.25%. The following plot also suggests that, the higher the average total monthly deposits a client has, the more likely this client would be to open an RRSP account with VanCity.

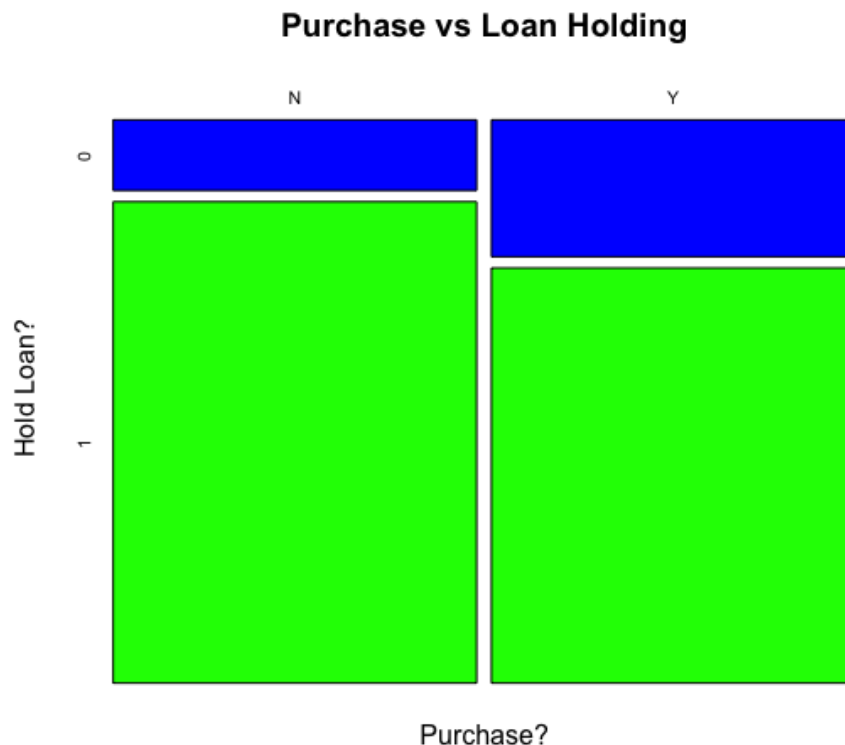


Additionally, the selection of this feature matches our first and third expectation in the mental model. First, assuming that the liability of a client is fixed, a higher averaged total deposit may indicate the client's strong income sources (of any type), meaning that he/she has greater cash surplus or positive network to participate in saving plans like RRSP. Additionally, stronger income sources become a strong indicator of their higher tax brackets. This is when they are in need of some plans like RRSP to help them lower taxable income and defer tax. Lastly, because a member's total average deposit is high, it's more likely for the client to be engaged with one of the financial advisors at VanCity to discuss the member's plan with their money. That is, it is

more likely to trigger an investment discussion where the clients would be suggested by VanCity sales teams to open an RRSP account if they did not have one.

2. *BALLOAN_IND*: whether a client holds a personal loan in the previous 12 months

We select it as an important predictor because it has a high magnitude of impact, a high significance level, and a large effect size. Our model shows that, if all other characteristics remain the same, the individuals with no personal loan are 73% more likely to become an RRSP purchaser. The following plot compares the distribution of purchasers across the loanholder group and the non-loanholder group. Although loanholders are the majority of both purchasers and non-purchasers, the non-loanholder group has a higher percentage of RRSP purchasers than the non-purchaser group.

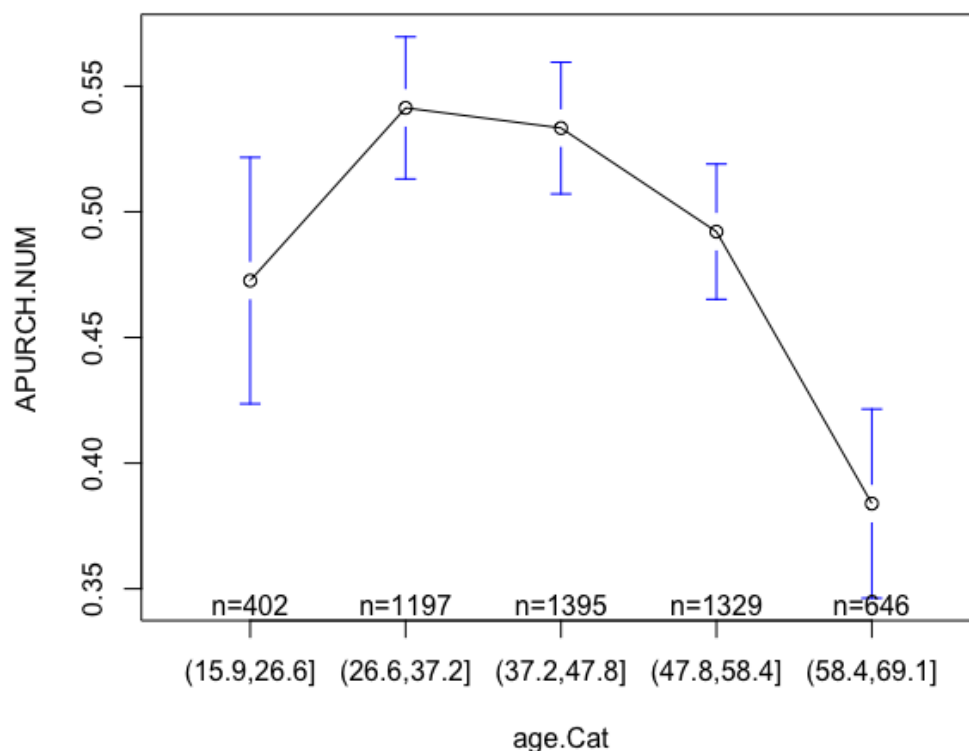


The predictor is chosen because it matches the capability (and the needs) consideration in our mental map. If a member does not hold a personal loan, he/she has less debt to pay off (or is in a debt-free status). The member may have more money compared to those who might have to pay off loans. The extra cash surplus increases the likelihood of investments like RRSP. We also need to consider why some members need a personal loan. Because the loans tend to have

higher interest rates than other borrowing products (such as credit lines), getting a personal loan is likely for specific personal uses including an emergency or necessary living expenses. Having a personal loan, different from other types of loan/mortgage, may indicate that the member is over budget or his/her financial situation is tight, in which case the likelihood of RRSP investment is low.

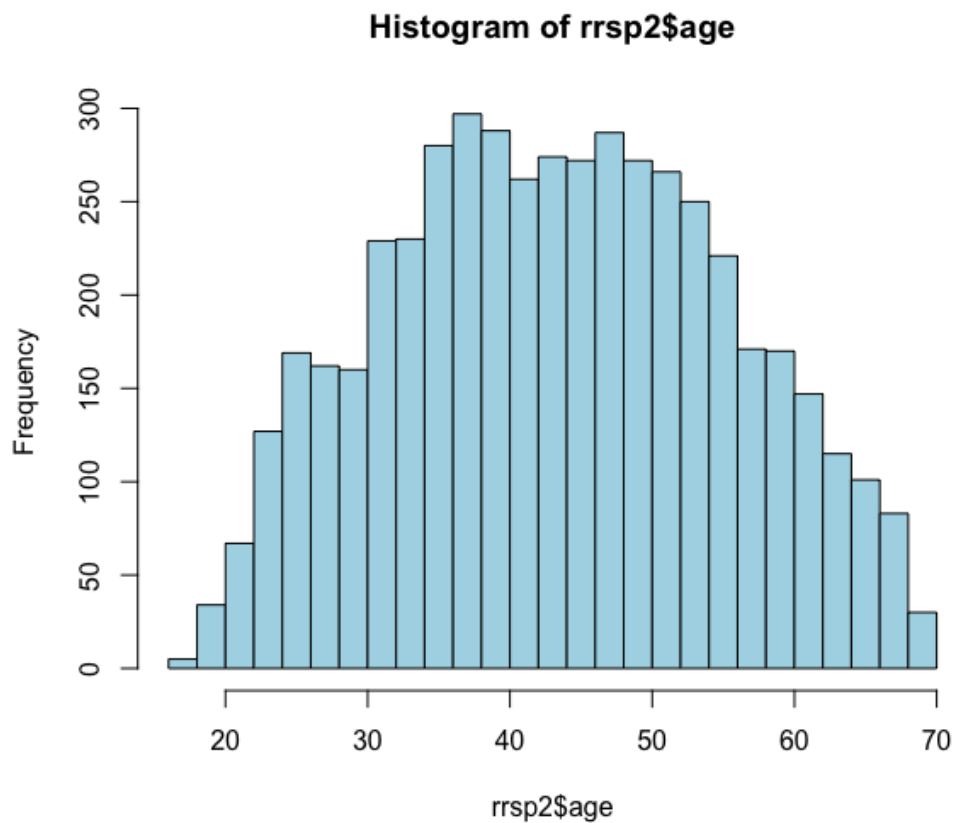
3. Age: the age of a client

Age is selected as an important predictor because it has a high significance level, a relatively high magnitude of impact, and a large effect size with a high confidence level. Our model shows that, for every increase of 1 in the age, the probability of purchase will decrease by 2.3%. We also consider the fact that government statistics suggest a strong pattern between the purchase propensity and age. The following plot shows a similar trend. As the age increases, the likelihood of RRSP purchases peaks in middle age and decreases rapidly.



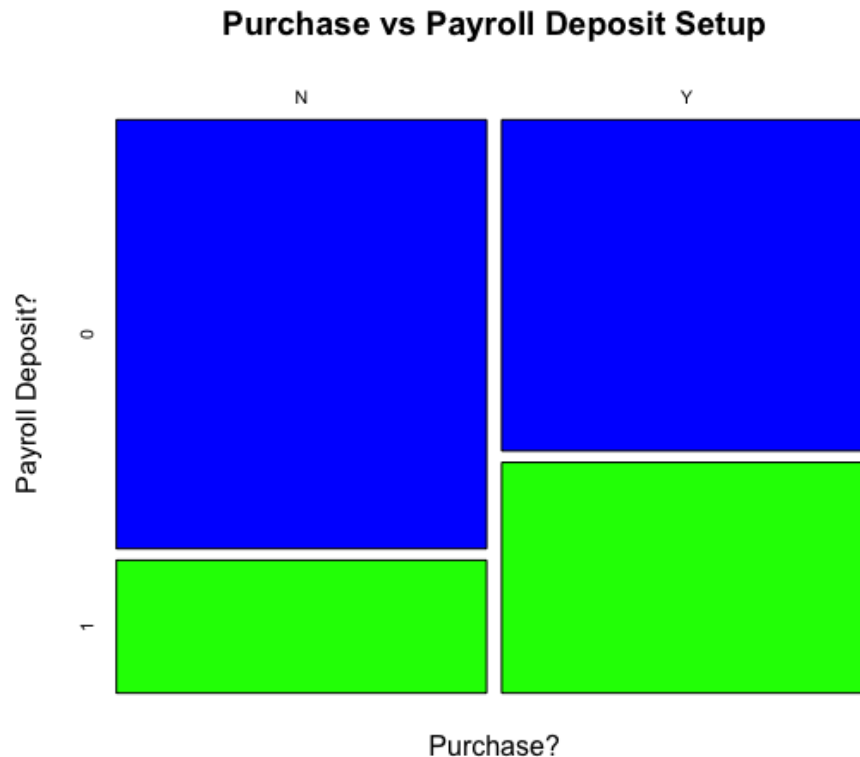
Choosing this feature matches the government statistics we have and the first conjecture in our mental map. Note that a majority of our clients (75%) are aged 35 years old or greater as shown in the histogram below, the model suggests an approximately linearly decreasing pattern from the middle age group. Moreover, the earning power of people in middle age is likely to be high

at this stage of life. They are the main target group for new RRSP purchases in the market. Because they earn more, they have higher tax brackets. So they are more likely to demand tax implications like an RRSP to defer tax (i.e., lower their taxable income and thus their tax payments). The other reason is that, because the maximum age for RRSP contributions is 71 years old, there are more younger clients purchasing and starting to invest into their new RRSP accounts than the older groups. Because the earlier they start to invest, the earlier they can take advantage of the effect of compounding.



4. Paydep: whether a member uses payroll deposits with VanCity

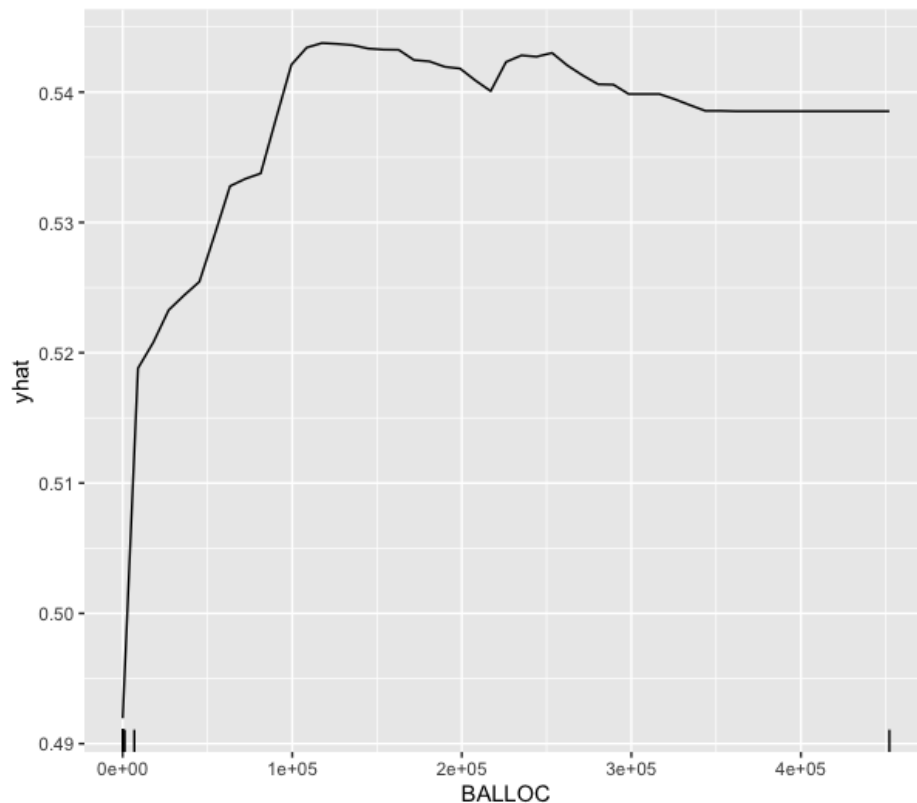
We choose payroll deposits as an important predictor because it has a high magnitude of impact, a high significance level, and a relatively large effect size. If a client has his/her payroll deposit set up with VanCity, holding all other conditions the same, he/she will be 38% more likely to purchase a new RRSP term. The following plot compares the distribution of purchasers across the setup group and the non-setup group. Although the majority of both purchasers and non-purchasers are those who do not use payroll deposit with VanCity, the setup group has a higher percentage of RRSP purchasers than the non-purchaser group.



The predictor corresponds to the intimacy and exposure consideration in our mental map. When a member sets up his/her payroll deposit with VanCity, the bank is likely to be his/her major bank. It thus becomes a strong indicator of intimacy and exposure. If VanCity is their major bank, the clients are more likely to receive RRSP campaigns from VanCity. Additionally, if VanCity is their major bank, clients are more likely to operate an RRSP account with VanCity instead of other competitors. Moreover, the RRSP campaigns are more likely to gain acceptance from clients with high intimacy.

5. Log.BALLOC: the average monthly line of credit balance over previous 12 months

We choose credit line balance as an important predictor because it has a relatively high significance level, magnitude of impact, and effect size. For those who have a credit line, every 1% increase in the credit line balance will increase the purchase possibility by 0.05%. The following plot shows that, as the balance of credit line increases, the probability of purchase is increasing at a decreasing rate.

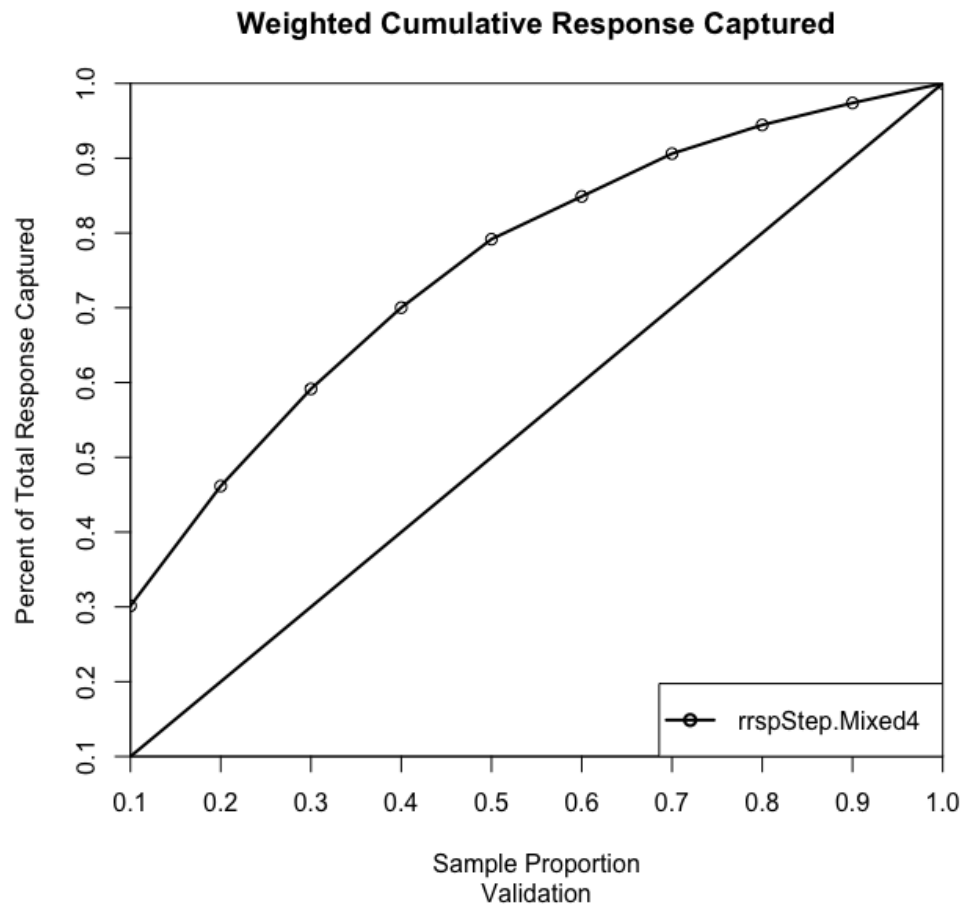


The positive relationship between the balance of credit lines and the purchase propensity is not as intuitive as the previous four features. As we discussed in the *BALLOAN_IND* part, a credit line is a type of debt which is supposed to decrease a client's cash surplus due to repayments. However, if we consider that a credit line has competitively low interest rates, it is not a main product for a bank to make profits but is more likely an incentive to encourage clients to borrow money with the bank to make investment (e.g., Home Equity Line of Credit). Additionally, a credit line is more likely to be approved to members who have a good relationship with VanCity and have a higher income. The better the relationship is and the higher the income is, the more the credit line will be approved to a client. While members do not necessarily use their credit lines, a higher balance may indicate that VanCity approved a larger credit line to a member based on an evaluation of the client's financial status and other profiles. The balance of credit lines may become an indicator of a member's income and relationship with the bank, which goes back to the capability and intimacy factors we discussed above, resulting in a high probability of purchase. Another possibility is that, having a credit line may indicate that a member's product category with VanCity is more diverse, meaning that he/she is more likely to be exposed to, accept, and purchase the RRSP product.

Question 3

TABLE 2A: Costs and Contribution (not the true proprietary figures!)

Contact cost (Mail and glossy brochure production cost) :	\$5.40
Number of potential contacts (members without an RRSP):	120,000
Estimated average contribution per RRSP purchase:	\$215.00



rrspStep.Mixed4 Lift Chart Point Values:

[1] 0.3014706 0.4616013 0.5915033 0.7001634 0.7916667 0.8488562 0.9060458 0.9444444
0.9738562 1.0000000

True Response Rate	Unit Cost	Unit Contribution				
2.20%	\$5.40	\$215.00				
Cumulative % contacted	Number contacted	% Cmtv Captured	Number Captured	Total Contact cost	Total Average Contribution	Total Profit
10	12,000	30.15%	796	\$64,800.00	\$171,114.71	\$106,314.71
20	24,000	46.16%	1,219	\$129,600.00	\$262,004.90	\$132,404.90
30	36,000	59.15%	1,562	\$194,400.00	\$335,737.27	\$141,337.27
40	48,000	70.02%	1,848	\$259,200.00	\$397,412.75	\$138,212.75
50	60,000	79.17%	2,090	\$324,000.00	\$449,350.02	\$125,350.02
60	72,000	84.89%	2,241	\$388,800.00	\$481,810.78	\$93,010.78
70	84,000	90.60%	2,392	\$453,600.00	\$514,271.60	\$60,671.60
80	96,000	94.44%	2,493	\$518,400.00	\$536,066.64	\$17,666.64
90	108,000	97.39%	2,571	\$583,200.00	\$552,760.78	-\$30,439.22
100	120,000	100%	2,640	\$648,000.00	\$567,600.00	-\$80,400.00

The above table uses the weighted cumulative lift chart of our best stepwise model to calculate, at each decile level, the number of total purchasers captured, the total contact cost, the total average contribution, and the total profit. We first obtain the number of purchasers captured at 100% by taking the product of the true response rate of 2.2% and the 120,000 potential contacts. The number captured at other decile levels are the product of their cumulative percentage contacted and the maximum possible number of purchasers. After we have all the values for column 4, we can easily calculate the contributions and profits.

Based on the results in the table, we recommend contacting the best 30% of the 120,000 potential members to maximize profits after contact costs. By contacting the best 36,000 clients ranked by our model, VanCity can expect to capture 1,562 purchasers. The cost of contacting 36,000 clients is \$194,400 and the expected contribution is \$335,737.27, giving us a maximal expected profit of \$141,337.27.