



Intuit Quickbooks Upgrade Campaign Analysis

Project Workflow

- * *Define business object*
- * *Data Quality Report (missing data, outliers, data errors)*
- * *Data Exploratory Analysis*
- * *Model Generation*
- * *Model Evaluation*
- * *Result Visualization*

1. Define business object

Increase Profit

Intuit Quickbooks has mailed 801821 customers an offer to upgrade to the latest version of the Quickbooks software in the wave-1 mailing, and 38487 users responded. Now, the company wants to mail this offer again to customers who did not respond to the wave-1 mailing.

Each mail piece costs 1.41\$ and the margin from each responder, excluding the mailing cost, is 60\$. "intuit75k.rds" file contains data on 75,000 (small) businesses that were selected randomly from the 801,821 that were sent the wave-1 mailing. In order to **achieve greater profit** through the wave-2 mailing, this project will use the available data to predict which businesses that did not respond to the wave-1 mailing, are most likely to respond to the wave-2 mailing.

Load Libraries

```
library(tidyverse)
library(radiant.data)
library(radiant)
library(readr)
library(glmnet)
library(caret)
library(pROC)
```

Loading the data

```
intuit75k <- readr::read_rds(file.path(radiant.data::find_dropbox(), "MGTA455-2019/data/intuit75k.rds"))
```

2. Data Quality Report

Data Description

Variable	Type	Description
id	integer	Small business customer ID
zip	character	5-Digit ZIP Code (00000 = unknown, 99999 = international ZIPs).
zip_bins	integer	Zip-code bins (20 approximately equal sized bins from lowest to highest zip code number)
sex	factor	"Female", "Male", or "Unknown."
bizflag	integer	Business Flag. Address contains a Business name (1=yes, 0=no or unknown).
numords	integer	Number of orders from Intuit Direct in the previous 36 months
dollars	numeric	Total \$ ordered from Intuit Direct in the previous 36 months
last	integer	Time (in months) since last order from Intuit Direct in the previous 36 months
sincepurch	integer	Time (in months) since original (not upgrade) Quickbooks purchase
version1	integer	Is 1 if the customer's current Quickbooks is version 1, 0 if version 2
owntaxprod	integer	Is 1 if the customer purchased tax software, 0 otherwise
upgraded	integer	Is 1 if customer upgraded from Quickbooks version 1 to version 2
res1	factor	Response to wave-1 mailing ("Yes" if responded else "No")
training	integer	70/30 split, 1 for training sample, 0 for validation sample

Data Type Transformation

```
intuit75k$zip_bins <- as.factor(intuit75k$zip_bins)
intuit75k$bizflag <- as.factor(intuit75k$bizflag)
intuit75k$version1 <- as.factor(intuit75k$version1)
intuit75k$owntaxprod <- as.factor(intuit75k$owntaxprod)
intuit75k$upgraded<- as.factor(intuit75k$upgraded)

categorical_variable <- intuit75k %>% select_if(is.factor)
numerical_variable <- intuit75k %>% select_if(is.numeric)
```

Data Summary - Numeric Variable

```
records_have_value = sapply(numerical_variable, function(x) sum(!is.na(x)))
populated = records_have_value/75000
unique_values = sapply(numerical_variable,n_distinct)
numer_of_zero_values = sapply(numerical_variable,function(x) sum(x==0))
mean = sapply(numerical_variable,mean)
std = sapply(numerical_variable,sd)
min = sapply(numerical_variable,min)
max = sapply(numerical_variable,max)
numeric_data_summary <- data.frame(records_have_value,populated,unique_values,numer_of_zero_values,mean,std,min,max)
numeric_data_summary
```

	records_have_value	populated	unique_values	numer_of_zero_values	mean	std	min	max
id	75000	1	75000	0	37500.50000	21	650.7794317	1 75000
numords	75000	1	5	0	2.07628		1.2413549	1 5
dollars	75000	1	1147	0	93.08713			

81.2059002	1	1149					
last			75000	1	36	0	15.83843
9.5390573	1	36					
sincepurch			75000	1	36	0	15.65717
10.0263119	1	36					
training			75000	1	2	22500	0.70000
0.4582606	0	1					

Data Summary - Categorical Variable

```
records_have_value = sapply(categorical_variable, function(x) sum(!is.na(x)))
populated = records_have_value/75000
unique_values = sapply(categorical_variable,n_distinct)
calculate_mode <- function(x) {
  uniqx <- unique(na.omit(x))
  uniqx[which.max(tabulate(match(x, uniqx)))]
}

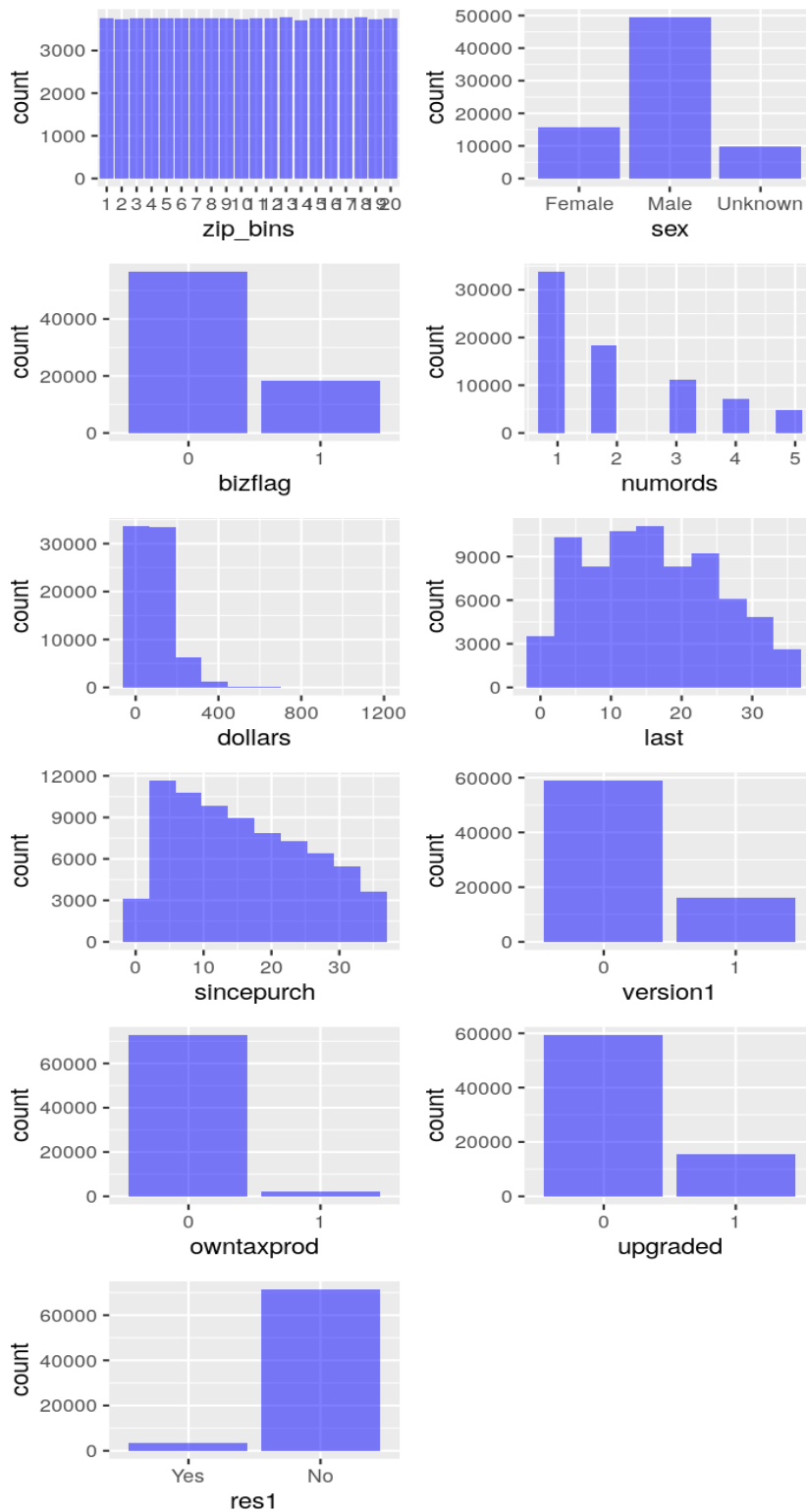
most_common_values = sapply(categorical_variable,calculate_mode)
categorical_data_summary <- data.frame(records_have_value,populated,unique_values,most_co
mmon_values)
categorical_data_summary
```

	records_have_value	populated	unique_values	most_common_values
zip_bins	75000	1	20	13
sex	75000	1	3	Male
bizflag	75000	1	2	0
version1	75000	1	2	0
owntaxprod	75000	1	2	0
upgraded	75000	1	2	0
res1	75000	1	2	No

3. Conduct EDA

Distributions of Variables

```
visualize(
  intuit75k,
  xvar = c(
    "zip_bins","sex", "bizflag", "numords", "dollars", "last", "sincepurch",
    "version1", "owntaxprod", "upgraded", "res1"
  ),
  type = "dist",
  custom = FALSE
)
```



Association between metric variables and response to wave-1

```
result <- compare_means(
  intuit75k,
  var1 = "res1",
  var2 = "numords"
)
summary(result, show = FALSE)
```

Pairwise mean comparisons (t-test)

Data : intuit75k
Variables : res1, numords
Samples : independent
Confidence: 0.95
Adjustment: None

res1	mean	n	sd	se	me
Yes	2.593	3,601	1.429	0.024	0.047
No	2.050	71,399	1.225	0.005	0.009

Null hyp.	Alt. hyp.	diff	p.value
Yes = No	Yes not equal to No	0.543	< .001 ***

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

```
result <- compare_means(  
  intuit75k,  
  var1 = "res1",  
  var2 = "dollars"  
)  
summary(result, show = FALSE)
```

Pairwise mean comparisons (t-test)

Data : intuit75k
Variables : res1, dollars
Samples : independent
Confidence: 0.95
Adjustment: None

res1	mean	n	sd	se	me
Yes	117.631	3,601	102.827	1.714	3.360
No	91.849	71,399	79.762	0.299	0.585

Null hyp.	Alt. hyp.	diff	p.value
Yes = No	Yes not equal to No	25.781	< .001 ***

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

```
result <- compare_means(  
  intuit75k,  
  var1 = "res1",  
  var2 = "last"  
)  
summary(result, show = FALSE)
```

Pairwise mean comparisons (t-test)

Data : intuit75k
Variables : res1, last
Samples : independent
Confidence: 0.95
Adjustment: None

res1	mean	n	sd	se	me
Yes	12.033	3,601	8.937	0.149	0.292
No	16.030	71,399	9.528	0.036	0.070

Null hyp.	Alt. hyp.	diff	p.value
Yes = No	Yes not equal to No	-3.998	< .001 ***

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

```
result <- compare_means(
  intuit75k,
  var1 = "res1",
  var2 = "sincepurch"
)
summary(result, show = FALSE)
```

Pairwise mean comparisons (t-test)

Data : intuit75k
 Variables : res1, sincepurch
 Samples : independent
 Confidence: 0.95
 Adjustment: None

	res1	mean	n	sd	se	me
Yes	19.180	3,601	9.936	0.166	0.325	
No	15.480	71,399	9.998	0.037	0.073	

Null hyp.	Alt. hyp.	diff	p.value
Yes = No	Yes not equal to No	3.7	< .001 ***

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

Correlation between variables

```
result <- correlation(
  intuit75k,
  vars = c(
    "zip_bins", "bizflag", "numords", "dollars", "last", "sincepurch",
    "version1", "owntaxprod", "upgraded", "res1"
  )
)
summary(result, covar = TRUE)
```

Correlation

Data : intuit75k
 Method : pearson
 Variables: zip_bins, bizflag, numords, dollars, last, sincepurch, version1, owntaxprod, upgraded, res1
 Null hyp.: variables x and y are not correlated
 Alt. hyp.: variables x and y are correlated

Correlation matrix:

	zip_bins	bizflag	numords	dollars	last	sincepurch	version1	owntaxprod	upgraded
bizflag	0.00								
numords	0.01	0.00							
dollars	0.01	0.00	0.59						
last	-0.00	-0.00	-0.13	-0.07					
sincepurch	-0.00	-0.01	0.00	0.00	-0.00				

version1	-0.00	-0.01	0.01	0.00	0.00	0.52			
owntaxprod	0.00	0.00	0.12	0.07	-0.02	-0.00	-0.08		
upgraded	-0.00	-0.00	-0.00	-0.00	-0.00	0.51	-0.27	0.09	
res1	0.06	-0.00	-0.09	-0.07	0.09	-0.08	-0.04	-0.03	-0.07

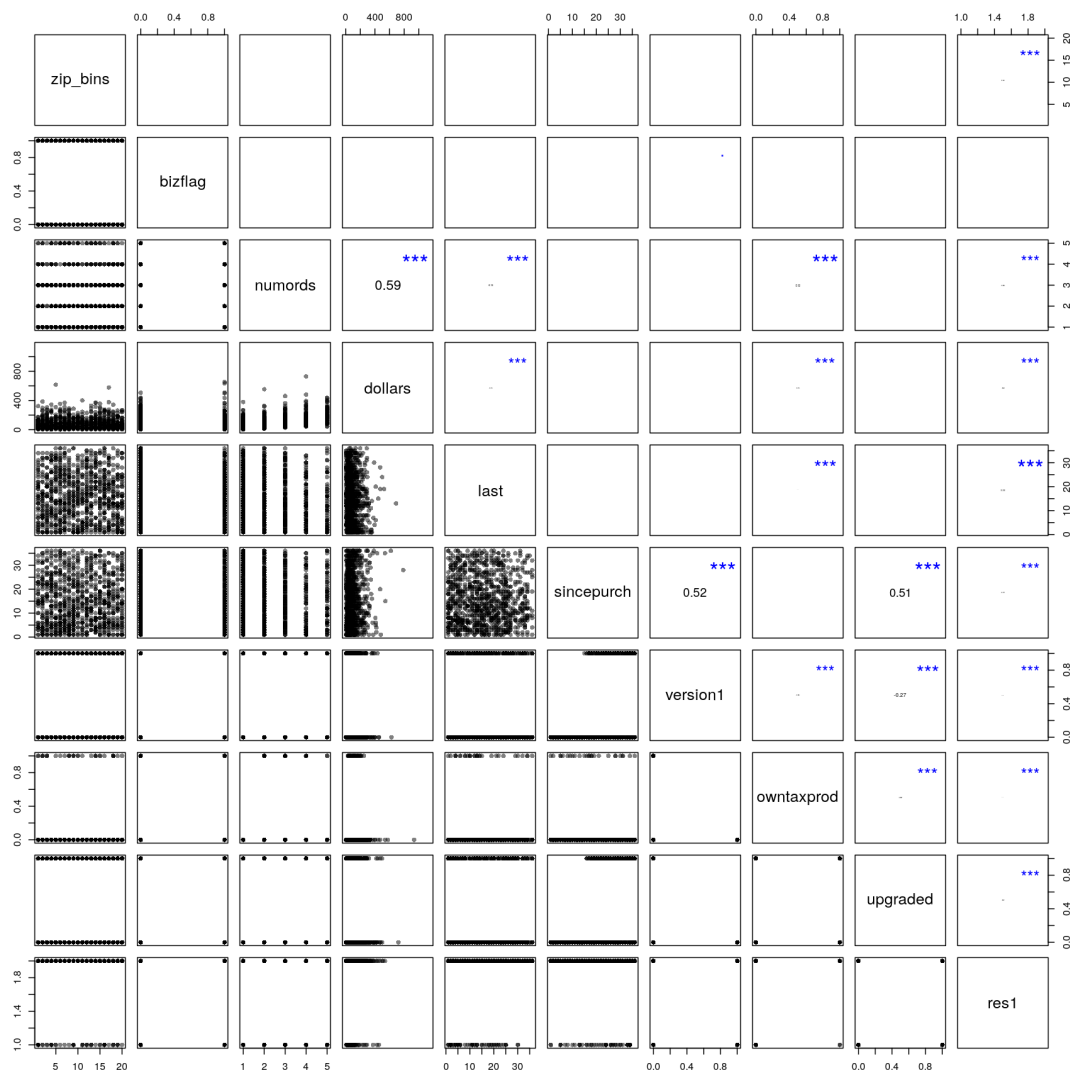
p.values:

	zip_bins	bizflag	numords	dollars	last	sincepurch	version1	owntaxprod	upgraded
bizflag	0.74								
numords	0.16	0.96							
dollars	0.11	0.58	0.00						
last	0.47	0.91	0.00	0.00					
sincepurch	0.65	0.15	0.63	0.82	0.75				
version1	0.38	0.10	0.12	0.54	0.58	0.00			
owntaxprod	0.57	0.21	0.00	0.00	0.00	0.75	0.00		
upgraded	0.97	0.40	0.45	0.70	0.96	0.00	0.00	0.00	
res1	0.00	0.83	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Covariance matrix:

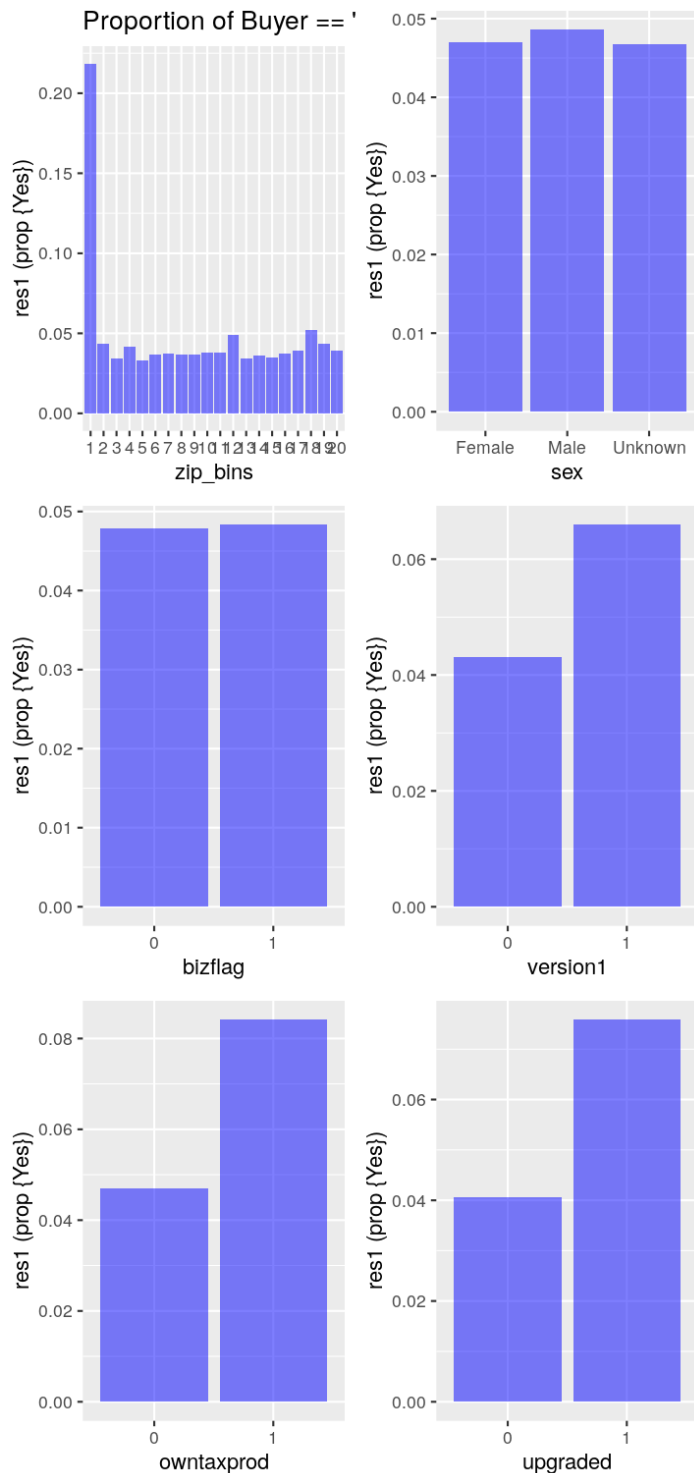
	zip_bins	bizflag	numords	dollars	last	sincepurch	version1	owntaxprod	upgrade
d									
bizflag	0.00								
numords	0.04	0.00							
dollars	2.73	0.07	59.04						
last	-0.15	-0.00	-1.54	-57.83					
sincepurch	-0.10	-0.02	0.02	0.66	-0.11				
version1	-0.01	-0.00	0.00	0.07	0.01	2.16			
owntaxprod	0.00	0.00	0.02	1.02	-0.03	-0.00	-0.01		
upgraded	-0.00	-0.00	-0.00	-0.05	-0.00	2.08	-0.04	0.01	
res1	0.08	-0.00	-0.02	-1.18	0.18	-0.17	-0.00	-0.00	-0.01

`plot(result, nrobs = 1000)`



Association between categorical variables and response to wave-1

```
visualize(
  intuit75k,
  xvar = c(
    "zip_bins", "sex", "bizflag", "version1", "owntaxprod", "upgraded"
  ),
  yvar = "res1",
  type = "bar",
  fun = "prop",
  labs = list(title = "Proportion of Buyer == 'Yes'"),
  custom = FALSE
)
```

According to the EDA part above, I noticed that the response rate in zip_bins = 1 is much higher than that of other zip_bins, therefore, I decided to further investigate what is going on in the first zip bin.

Which zip code is skewing bin 1 so high for proportion of buyers?

```
intuit75k %>%
  filter(res1 == "Yes" & zip_bins == "1") %>%
  count(zip) %>%
  arrange(desc(n)) %>%
  slice(1:10)
```

```
# A tibble: 10 x 2
  zip      n
  <chr> <int>
1 00801  688
2 00804   64
3 00000    5
4 01923    4
5 01890    3
6 01504    2
7 01752    2
8 01754    2
9 01863    2
10 01950    2
```

What do some of the response rates look like for the common zip codes that included in dataset?

```
intuit75k %>%
  group_by(zip) %>%
  summarize(total = sum(n()),
            num_resp = sum(res1 == "Yes"),
            no_resp = total - num_resp,
            resp_rate = num_resp / total) %>%
  arrange(desc(total)) %>%
  slice(1:10)
```

```
# A tibble: 10 x 5
  zip      total num_resp no_resp resp_rate
  <chr> <int>      <int>   <int>   <dbl>
1 00801  1668         688     980    0.412
2 99999   648          19     629    0.0293
3 00804   186          64     122    0.344
4 00000   137           5     132    0.0365
5 92714    64           7      57    0.109
6 94087    63           3      60    0.0476
7 10021    59           1      58    0.0169
8 10022    59           5      54    0.0847
9 94025    56           3      53    0.0536
10 94596    55           4      51    0.0727
```

From the analysis above, it is obvious that the response rate in zip-code 00801 and 00804 is way higher than that of other zip-codes, therefore, creating features to denote this might be helpful to model building. Hence, I create a new variable to demonstrate this called zip_801 and zip_804.

```
intuit75k <- mutate(intuit75k, zip_801 = ifelse(zip == "00801" , 1, 0), zip_804 = ifelse(
zip == "00804" , 1, 0))
```

4. Model Generation

Define Train/Test Data

```
training <- intuit75k %>% filter(training == 1)
testing <- intuit75k %>% filter(training == 0)
```

Models:

Estimate a logistic regression model with all original variables

```
resultl1 <- logistic(  
  training,  
  rvar = "res1",  
  evar = c(  
    "zip_bins", "sex", "bizflag", "numords", "dollars", "last",  
    "sincepurch", "version1", "owntaxprod", "upgraded"  
  ),  
  lev = "Yes"  
)  
summary(resultl1, sum_check = "odds")
```

Logistic regression (GLM)

Data : training

Response variable : res1

Level : Yes in res1

Explanatory variables: zip_bins, sex, bizflag, numords, dollars, last, sincepurch, version1, owntaxprod, upgraded

Null hyp.: there is no effect of x on res1

Alt. hyp.: there is an effect of x on res1

		OR	coefficient	std.error	z.value	p.value	
(Intercept)		-3.761	0.211	-17.815	< .001	***	
zip_bins 2	0.148	-1.910	0.110	-17.297	< .001	***	
zip_bins 3	0.118	-2.141	0.121	-17.710	< .001	***	
zip_bins 4	0.136	-1.996	0.112	-17.811	< .001	***	
zip_bins 5	0.116	-2.152	0.120	-17.894	< .001	***	
zip_bins 6	0.126	-2.068	0.115	-18.043	< .001	***	
zip_bins 7	0.121	-2.110	0.117	-17.970	< .001	***	
zip_bins 8	0.131	-2.034	0.114	-17.844	< .001	***	
zip_bins 9	0.124	-2.086	0.117	-17.856	< .001	***	
zip_bins 10	0.121	-2.110	0.118	-17.919	< .001	***	
zip_bins 11	0.128	-2.053	0.116	-17.642	< .001	***	
zip_bins 12	0.172	-1.763	0.104	-16.871	< .001	***	
zip_bins 13	0.113	-2.177	0.121	-17.940	< .001	***	
zip_bins 14	0.131	-2.031	0.115	-17.718	< .001	***	
zip_bins 15	0.118	-2.135	0.119	-17.892	< .001	***	
zip_bins 16	0.130	-2.042	0.114	-17.896	< .001	***	
zip_bins 17	0.130	-2.038	0.114	-17.848	< .001	***	
zip_bins 18	0.183	-1.698	0.101	-16.745	< .001	***	
zip_bins 19	0.143	-1.947	0.111	-17.500	< .001	***	
zip_bins 20	0.124	-2.084	0.116	-17.973	< .001	***	
sex Male	0.975	-0.025	0.053	-0.472	0.637		
sex Unknown	0.964	-0.037	0.076	-0.485	0.628		
bizflag	1.036	0.036	0.049	0.728	0.466		
numords	1.259	0.230	0.019	12.047	< .001	***	
dollars	1.001	0.001	0.000	4.009	< .001	***	
last	0.957	-0.044	0.002	-18.064	< .001	***	
sincepurch	1.002	0.002	0.004	0.467	0.641		
version1	2.113	0.748	0.087	8.583	< .001	***	
owntaxprod	1.356	0.304	0.103	2.959	0.003	**	
upgraded	2.616	0.962	0.086	11.201	< .001	***	

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

Pseudo R-squared: 0.114

Log-likelihood: -8899.946, AIC: 17859.893, BIC: 18125.95

Chi-squared: 2289.751 df(29), p.value < .001

Nr obs: 52,500

	odds	ratio	2.5%	97.5%
zip_bins 2	0.148	0.119	0.184	
zip_bins 3	0.118	0.093	0.149	
zip_bins 4	0.136	0.109	0.169	
zip_bins 5	0.116	0.092	0.147	
zip_bins 6	0.126	0.101	0.158	
zip_bins 7	0.121	0.096	0.153	
zip_bins 8	0.131	0.105	0.163	
zip_bins 9	0.124	0.099	0.156	
zip_bins 10	0.121	0.096	0.153	
zip_bins 11	0.128	0.102	0.161	
zip_bins 12	0.172	0.140	0.211	
zip_bins 13	0.113	0.089	0.144	
zip_bins 14	0.131	0.105	0.164	
zip_bins 15	0.118	0.094	0.149	
zip_bins 16	0.130	0.104	0.162	
zip_bins 17	0.130	0.104	0.163	
zip_bins 18	0.183	0.150	0.223	
zip_bins 19	0.143	0.115	0.178	
zip_bins 20	0.124	0.099	0.156	
sex Male	0.975	0.879	1.082	
sex Unknown	0.964	0.831	1.118	
bizflag	1.036	0.941	1.141	
numords	1.259	1.212	1.307	
dollars	1.001	1.001	1.002	
last	0.957	0.953	0.962	
sincepurch	1.002	0.994	1.010	
version1	2.113	1.781	2.506	
owntaxprod	1.356	1.108	1.659	
upgraded	2.616	2.211	3.096	

Estimate a logistic regression model with new feature zip_801 and zip_804

```
resultl2 <- logistic(  
  training,  
  rvar = "res1",  
  evar = c(  
    "zip_bins", "sex", "bizflag", "numords", "dollars", "last",  
    "sincepurch", "version1", "owntaxprod", "upgraded", "zip_801", "zip_804"  
  ),  
  lev = "Yes",  
)  
summary(resultl2, sum_check = "odds")
```

Logistic regression (GLM)

Data : training

Response variable : res1

Level : Yes in res1
 Explanatory variables: zip_bins, sex, bizflag, numords, dollars, last, sincepurch, version1, owntaxprod, upgraded, zip_801, zip_804
 Null hyp.: there is no effect of x on res1
 Alt. hyp.: there is an effect of x on res1

	OR	coefficient	std.error	z.value	p.value	
(Intercept)		-6.119	0.261	-23.403	< .001	***
zip_bins 2	1.300	0.262	0.180	1.457	0.145	
zip_bins 3	1.031	0.031	0.187	0.165	0.869	
zip_bins 4	1.189	0.173	0.181	0.956	0.339	
zip_bins 5	1.020	0.020	0.186	0.109	0.913	
zip_bins 6	1.107	0.102	0.183	0.557	0.578	
zip_bins 7	1.063	0.061	0.184	0.331	0.741	
zip_bins 8	1.147	0.137	0.182	0.751	0.453	
zip_bins 9	1.088	0.085	0.184	0.461	0.645	
zip_bins 10	1.063	0.062	0.185	0.334	0.739	
zip_bins 11	1.126	0.119	0.184	0.646	0.518	
zip_bins 12	1.505	0.409	0.176	2.319	0.020	*
zip_bins 13	0.995	-0.005	0.187	-0.024	0.981	
zip_bins 14	1.150	0.140	0.183	0.765	0.444	
zip_bins 15	1.037	0.036	0.186	0.195	0.845	
zip_bins 16	1.137	0.129	0.182	0.706	0.480	
zip_bins 17	1.141	0.132	0.182	0.723	0.470	
zip_bins 18	1.606	0.474	0.175	2.714	0.007	**
zip_bins 19	1.251	0.224	0.181	1.242	0.214	
zip_bins 20	1.090	0.086	0.183	0.471	0.638	
sex Male	0.988	-0.012	0.054	-0.223	0.823	
sex Unknown	0.957	-0.044	0.077	-0.572	0.568	
bizflag	1.047	0.046	0.050	0.917	0.359	
numords	1.281	0.248	0.019	12.707	< .001	***
dollars	1.001	0.001	0.000	4.145	< .001	***
last	0.956	-0.045	0.002	-18.299	< .001	***
sincepurch	1.002	0.002	0.004	0.519	0.604	
version1	2.184	0.781	0.089	8.759	< .001	***
owntaxprod	1.384	0.325	0.105	3.112	0.002	**
upgraded	2.749	1.011	0.088	11.511	< .001	***
zip_801 1	25.333	3.232	0.164	19.746	< .001	***
zip_804 1	18.398	2.912	0.249	11.709	< .001	***

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

Pseudo R-squared: 0.148
 Log-likelihood: -8560.817, AIC: 17185.633, BIC: 17469.428
 Chi-squared: 2968.011 df(31), p.value < .001
 Nr obs: 52,500

	odds	ratio	2.5%	97.5%
zip_bins 2	1.300	0.914	1.850	
zip_bins 3	1.031	0.715	1.487	
zip_bins 4	1.189	0.834	1.695	
zip_bins 5	1.020	0.708	1.470	
zip_bins 6	1.107	0.774	1.583	
zip_bins 7	1.063	0.741	1.526	
zip_bins 8	1.147	0.802	1.639	

zip_bins 9	1.088	0.759	1.561
zip_bins 10	1.063	0.741	1.527
zip_bins 11	1.126	0.786	1.614
zip_bins 12	1.505	1.065	2.127
zip_bins 13	0.995	0.690	1.436
zip_bins 14	1.150	0.804	1.645
zip_bins 15	1.037	0.721	1.492
zip_bins 16	1.137	0.796	1.626
zip_bins 17	1.141	0.798	1.631
zip_bins 18	1.606	1.141	2.262
zip_bins 19	1.251	0.879	1.783
zip_bins 20	1.090	0.761	1.562
sex Male	0.988	0.888	1.099
sex Unknown	0.957	0.822	1.114
bizflag	1.047	0.949	1.155
numords	1.281	1.233	1.331
dollars	1.001	1.001	1.002
last	0.956	0.951	0.960
sincepurch	1.002	0.994	1.010
version1	2.184	1.834	2.601
owntaxprod	1.384	1.128	1.699
upgraded	2.749	2.314	3.266
zip_801 1	25.333	18.381	34.915
zip_804 1	18.398	11.300	29.956

Estimate a Neural Network model with all variables with one hidden layer

```
resultn1 <- nn(
  training,
  rvar = "res1",
  evar = c(
    "zip_bins", "sex", "bizflag", "numords", "dollars", "last",
    "sincepurch", "version1", "owntaxprod", "upgraded", "zip_801", "zip_804"
  ),
  lev = "Yes",
  seed = 1234
)
summary(resultn1, prn = TRUE)
```

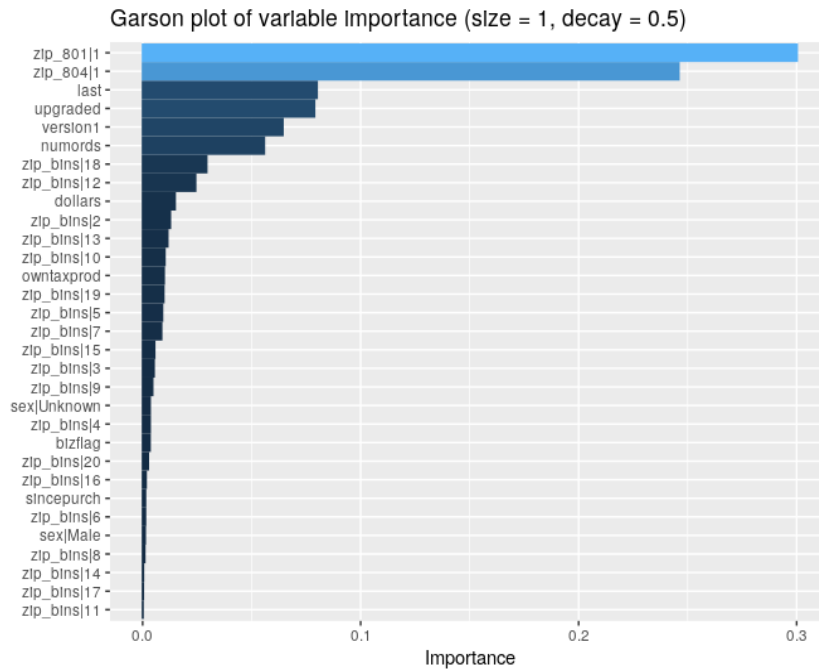
```
Neural Network
Activation function : Logistic (classification)
Data                : training
Response variable   : res1
Level               : Yes in res1
Explanatory variables: zip_bins, sex, bizflag, numords, dollars, last, sincepurch, versio
n1, owntaxprod, upgraded, zip_801, zip_804
Network size        : 1
Parameter decay     : 0.5
Seed                : 1234
Network             : 31-1-1 with 34 weights
Nr obs              : 52,500
Weights             :
  b->h1  i1->h1  i2->h1  i3->h1  i4->h1  i5->h1  i6->h1  i7->h1  i8->h1  i9->h1  i10->h1
i11->h1
      0.93   -0.11    0.05   -0.03    0.08    0.01    0.08    0.01    0.04    0.09    0.00
```

```

-0.21
i12->h1 i13->h1 i14->h1 i15->h1 i16->h1 i17->h1 i18->h1 i19->h1 i20->h1 i21->h1 i22->h1
i23->h1
0.10 0.00 0.05 -0.01 0.00 -0.26 -0.08 0.02 -0.01 0.03 -0.03
-0.49
i24->h1 i25->h1 i26->h1 i27->h1 i28->h1 i29->h1 i30->h1 i31->h1
-0.13 0.70 -0.01 -0.56 -0.09 -0.69 -2.61 -2.14
b->o h1->o
0.60 -5.84

```

```
plot(resultn1, plots = "garson", custom = FALSE)
```



Estimate a Neural Network model with all variables with two hidden layers

```

resultn2 <- nn(
  training,
  rvar = "res1",
  evar = c(
    "zip_bins", "sex", "bizflag", "numords", "dollars", "last",
    "sincepurch", "version1", "owntaxprod", "upgraded", "zip_801", "zip_804"
  ),
  lev = "Yes",
  size = 2,
  seed = 1234
)
summary(resultn2, prn = TRUE)

```

```

Neural Network
Activation function : Logistic (classification)
Data               : training
Response variable  : res1
Level              : Yes in res1
Explanatory variables: zip_bins, sex, bizflag, numords, dollars, last, sincepurch, version1, owntaxprod, upgraded, zip_801, zip_804
Network size       : 2

```

```

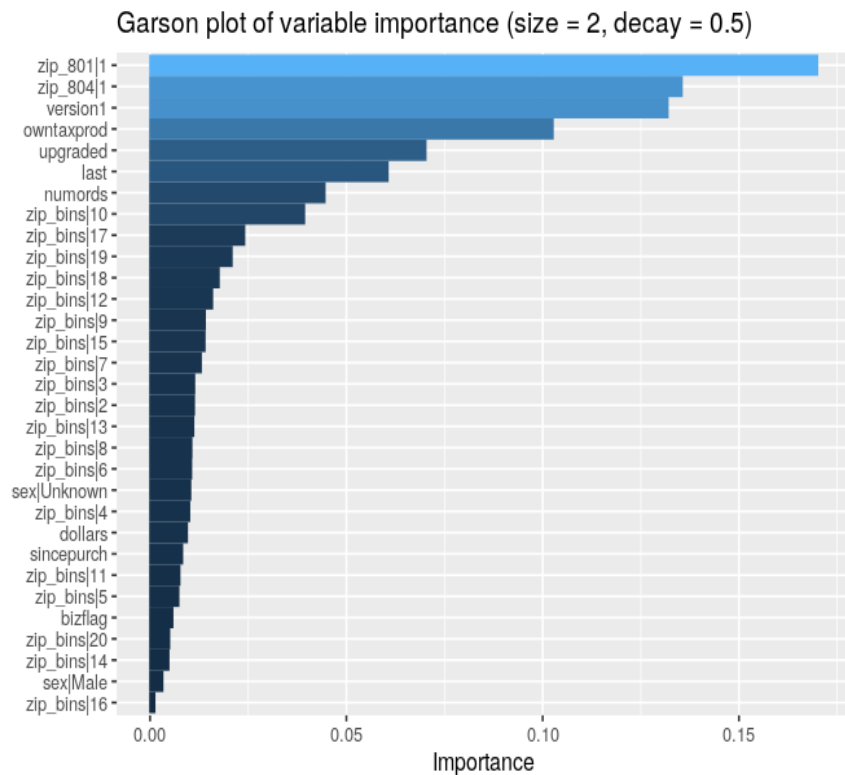
Parameter decay      : 0.5
Seed                 : 1234
Network              : 31-2-1 with 67 weights
Nr obs               : 52,500
Weights              :
  b->h1  i1->h1  i2->h1  i3->h1  i4->h1  i5->h1  i6->h1  i7->h1  i8->h1  i9->h1  i10->h1
i11->h1
  -1.00  -0.34   0.34  -0.18   0.02  -0.18  -0.13  -0.15  -0.21  -0.62   0.19
  -0.30
  i12->h1 i13->h1 i14->h1 i15->h1 i16->h1 i17->h1 i18->h1 i19->h1 i20->h1 i21->h1 i22->h1
i23->h1
   0.13   0.10   0.38  -0.01  -0.45  -0.25   0.36   0.15   0.06   0.23  -0.14
  -1.01
  i24->h1 i25->h1 i26->h1 i27->h1 i28->h1 i29->h1 i30->h1 i31->h1
  -0.12   1.33  -0.19   1.94   1.90  -1.37  -1.99  -1.33
  b->h2  i1->h2  i2->h2  i3->h2  i4->h2  i5->h2  i6->h2  i7->h2  i8->h2  i9->h2  i10->h2
i11->h2
   1.30  -0.02   0.01   0.12   0.17   0.13   0.23   0.15   0.19   0.52  -0.04
  -0.17
  i12->h2 i13->h2 i14->h2 i15->h2 i16->h2 i17->h2 i18->h2 i19->h2 i20->h2 i21->h2 i22->h2
i23->h2
   0.18   0.04  -0.05   0.02   0.26  -0.26  -0.25  -0.01  -0.03  -0.08   0.04
  -0.34
  i24->h2 i25->h2 i26->h2 i27->h2 i28->h2 i29->h2 i30->h2 i31->h2
  -0.15   0.50   0.06  -1.86  -1.13  -0.72  -2.81  -2.45
  b->o  h1->o  h2->o
   0.99 -3.12 -4.65

```

```

plot(resultn2, plots = "garson", custom = FALSE)

```



Estimate a Neural Network model with all variables with three hidden layers

```
resultn3 <- nn(
  training,
  rvar = "res1",
  evar = c(
    "zip_bins", "sex", "bizflag", "numords", "dollars", "last",
    "sincepurch", "version1", "owntaxprod", "upgraded", "zip_801", "zip_804"
  ),
  lev = "Yes",
  size = 3,
  seed = 1234
)
summary(resultn3, prn = TRUE)
```

Neural Network

Activation function : Logistic (classification)

Data : training

Response variable : res1

Level : Yes in res1

Explanatory variables: zip_bins, sex, bizflag, numords, dollars, last, sincepurch, version1, owntaxprod, upgraded, zip_801, zip_804

Network size : 3

Parameter decay : 0.5

Seed : 1234

Network : 31-3-1 with 100 weights

Nr obs : 52,500

Weights :

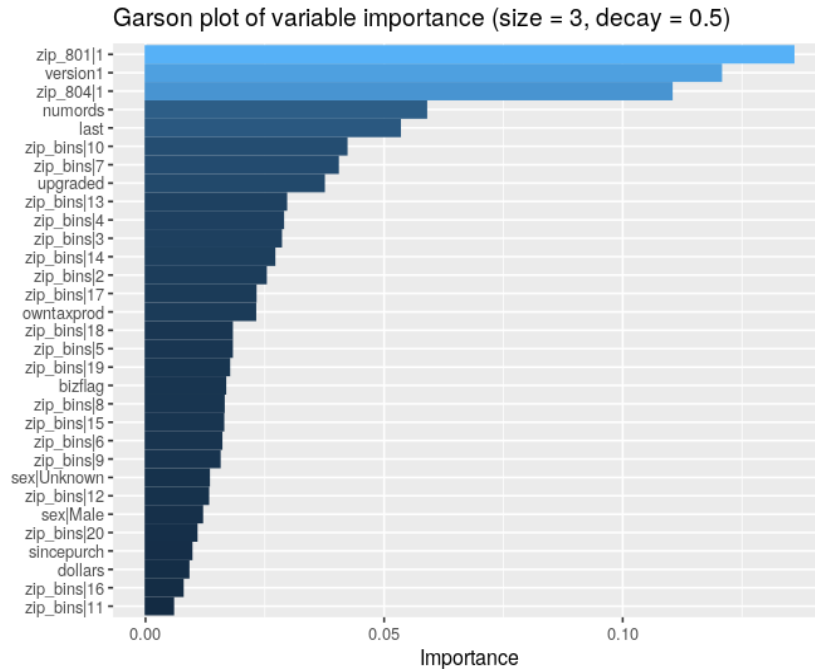
b->h1	i1->h1	i2->h1	i3->h1	i4->h1	i5->h1	i6->h1	i7->h1	i8->h1	i9->h1	i10->h1	
i11->h1	1.44	-0.43	0.15	0.53	-0.03	-0.33	0.85	0.25	0.30	0.90	-0.11
-0.44											
i12->h1	i13->h1	i14->h1	i15->h1	i16->h1	i17->h1	i18->h1	i19->h1	i20->h1	i21->h1	i22->h1	
i23->h1	0.17	0.36	-0.22	-0.12	0.38	-0.15	0.24	-0.12	0.23	0.06	0.28
-1.00											
i24->h1	i25->h1	i26->h1	i27->h1	i28->h1	i29->h1	i30->h1	i31->h1				
-0.11	0.49	0.14	-0.93	0.35	-0.05	-2.49	-2.23				
b->h2	i1->h2	i2->h2	i3->h2	i4->h2	i5->h2	i6->h2	i7->h2	i8->h2	i9->h2	i10->h2	
i11->h2	-1.15	-0.28	0.56	-0.24	0.49	-0.01	-0.67	0.07	-0.26	-0.40	0.10
-0.09											
i12->h2	i13->h2	i14->h2	i15->h2	i16->h2	i17->h2	i18->h2	i19->h2	i20->h2	i21->h2	i22->h2	
i23->h2	0.60	0.20	0.27	-0.06	-0.55	-0.27	-0.25	-0.10	-0.07	0.26	-0.19
-0.87											
i24->h2	i25->h2	i26->h2	i27->h2	i28->h2	i29->h2	i30->h2	i31->h2				
0.05	1.61	-0.01	2.07	-0.06	-0.56	-1.55	-0.67				
b->h3	i1->h3	i2->h3	i3->h3	i4->h3	i5->h3	i6->h3	i7->h3	i8->h3	i9->h3	i10->h3	
i11->h3	-0.31	0.31	-0.41	-0.39	-0.19	0.30	-0.15	-0.33	-0.09	-0.42	0.03
0.02											
i12->h3	i13->h3	i14->h3	i15->h3	i16->h3	i17->h3	i18->h3	i19->h3	i20->h3	i21->h3	i22->h3	
i23->h3	-0.40	-0.51	0.17	0.12	-0.03	-0.29	-0.21	0.20	-0.18	-0.20	-0.20

```

0.52
i24->h3 i25->h3 i26->h3 i27->h3 i28->h3 i29->h3 i30->h3 i31->h3
-0.20 -0.08 -0.22 -1.76 -0.49 -0.83 -1.47 -1.56
b->o h1->o h2->o h3->o
1.70 -3.64 -3.16 -3.41

```

```
plot(resultn3, plots = "garson", custom = FALSE)
```



From the three garson plots above, we can noticed that the importance of variable version1 changed a lot from neural network model with one hidden layer to the model with three hidden layers, which shows that there are non-linearity or interaction effects between version1 and the response variable. Next,let us go deeper into the effect of version1 on res1.

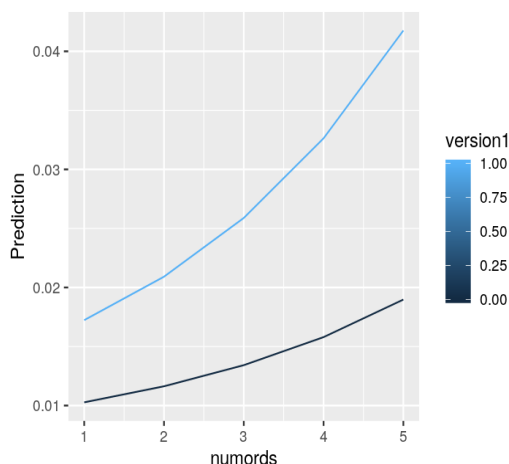
Explore Interaction Effects

Explore the effect of version1 on response rate under different number of orders while holding all other variables constant.

```

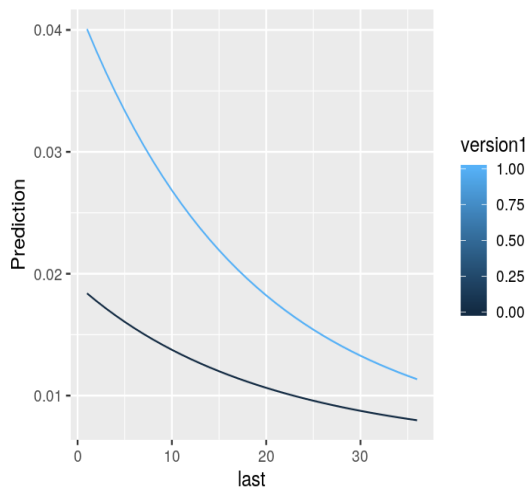
pred <- predict(resultn1, pred_cmd = "numords = 1:5,version1 = c(0,1)")
plot(pred, xvar = "numords",color = "version1")

```



It is shown in the above chart that there the effect of version1 on res1 is influenced by the value of numords, which suggests that there is interaction between version1 and numords.

```
pred <- predict(resultn1, pred_cmd = "last = 1:36,version1 = c(0,1)")
plot(pred, xvar = "last",color = "version1")
```



It is shown in the above chart that there the effect of version1 on res1 is influenced by the value of last, which suggests that there is interaction between version1 and last.

Add Interaction Effect to Logistic Model

```
resultl3 <- logistic(
  training,
  rvar = "res1",
  evar = c(
    "zip_bins","sex", "bizflag", "numords", "dollars", "last",
    "sincepurch", "version1", "owntaxprod", "upgraded","zip_801","zip_804"
  ),
  lev = "Yes",
  int = c("version1:numords","version1:last")
)
```

```
summary(resultl3)
```

Logistic regression (GLM)

Data : training

Response variable : res1

Level : Yes in res1

Explanatory variables: zip_bins, sex, bizflag, numords, dollars, last, sincepurch, version1, owntaxprod, upgraded, zip_801, zip_804

Null hyp.: there is no effect of x on res1

Alt. hyp.: there is an effect of x on res1

	OR	coefficient	std.error	z.value	p.value
(Intercept)		-5.689	0.301	-18.904	< .001 ***
zip_bins 2	1.296	0.259	0.181	1.436	0.151
zip_bins 3	1.034	0.034	0.187	0.181	0.857
zip_bins 4	1.180	0.166	0.182	0.912	0.362

zip_bins 5	1.015	0.015	0.187	0.078	0.938	
zip_bins 6	1.099	0.094	0.183	0.514	0.607	
zip_bins 7	1.055	0.053	0.185	0.287	0.774	
zip_bins 8	1.118	0.111	0.183	0.608	0.543	
zip_bins 9	1.074	0.072	0.185	0.388	0.698	
zip_bins 10	1.060	0.058	0.185	0.314	0.753	
zip_bins 11	1.122	0.116	0.184	0.627	0.531	
zip_bins 12	1.504	0.408	0.177	2.307	0.021	*
zip_bins 13	0.985	-0.015	0.188	-0.081	0.936	
zip_bins 14	1.129	0.122	0.183	0.664	0.507	
zip_bins 15	1.019	0.019	0.186	0.102	0.919	
zip_bins 16	1.123	0.116	0.183	0.633	0.527	
zip_bins 17	1.119	0.112	0.183	0.614	0.539	
zip_bins 18	1.604	0.472	0.175	2.695	0.007	**
zip_bins 19	1.243	0.217	0.181	1.201	0.230	
zip_bins 20	1.087	0.083	0.184	0.451	0.652	
sex Male	0.992	-0.008	0.055	-0.154	0.878	
sex Unknown	0.964	-0.037	0.078	-0.471	0.638	
bizflag	1.048	0.046	0.050	0.923	0.356	
numords	0.832	-0.184	0.050	-3.682	< .001	***
dollars	1.001	0.001	0.000	4.208	< .001	***
last	1.007	0.007	0.008	0.864	0.388	
sincepurch	1.002	0.002	0.004	0.531	0.595	
version1	1.474	0.388	0.151	2.572	0.010	*
owntaxprod	1.496	0.403	0.105	3.849	< .001	***
upgraded	2.719	1.000	0.088	11.371	< .001	***
zip_801 1	25.908	3.255	0.164	19.832	< .001	***
zip_804 1	18.400	2.912	0.250	11.661	< .001	***
numords:version1	1.397	0.334	0.035	9.442	< .001	***
last:version1	0.959	-0.042	0.006	-7.089	< .001	***

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

Pseudo R-squared: 0.156

Log-likelihood: -8479.762, AIC: 17027.524, BIC: 17329.055

Chi-squared: 3130.12 df(33), p.value < .001

Nr obs: 52,500

Prediction Using Different Models

Set the break even rate

```
mail_cost <- 1.41
margin_sales <- 60
breakeven <- mail_cost / margin_sales
```

Prediction

```
## Logistic Model
predl3 <- predict(resultl3, pred_data = testing)
testing <- store(testing, predl3, name = "purch_prob13")

## Neural Network Model
predn1 <- predict(resultn1, pred_data = testing)
```

```

testing <- store(testing, predn1, name = "purch_probn1")
predn2 <- predict(resultn2, pred_data = testing)
testing <- store(testing, predn2, name = "purch_probn2")
predn3 <- predict(resultn3, pred_data = testing)
testing <- store(testing, predn3, name = "purch_probn3")

## Ensemble Logistic and Neural Network Model
testing$purch_ensemble <- (testing$purch_probn1 + testing$purch_probn2 + testing$purch_probn3)/4

## Create Mailto Variable)
testing <- testing %>%
  mutate(mailto_logit = purch_probn1/2 > breakeven,
         mailto_n1 = purch_probn2/2 > breakeven,
         mailto_n2 = purch_probn3/2 > breakeven,
         mailto_ensemble = purch_ensemble/2 > breakeven)

```

5. Model Comparison

```

prediction_result <- tibble::tibble(
  Logistic = as.integer(testing$mailto_logit),
  Neural_network1 = as.integer(testing$mailto_n1),
  Neural_network2 = as.integer(testing$mailto_n2),
  Neural_network3 = as.integer(testing$mailto_n3),
  Ensemble = as.integer(testing$mailto_ensemble),
  res1 = testing$res1)

compare_result <- confusion(
  prediction_result,
  pred = c("Logistic", "Neural_network1", "Neural_network2", "Neural_network3", "Ensemble"),
  rvar = "res1",
  lev = "Yes",
  cost = 1.41,
  margin = 60
)
summary(compare_result)

```

Confusion matrix

Data : prediction_result

Results for: All

Predictors : Logistic, Neural_network1, Neural_network2, Neural_network3, Ensemble

Response : res1

Level : Yes in res1

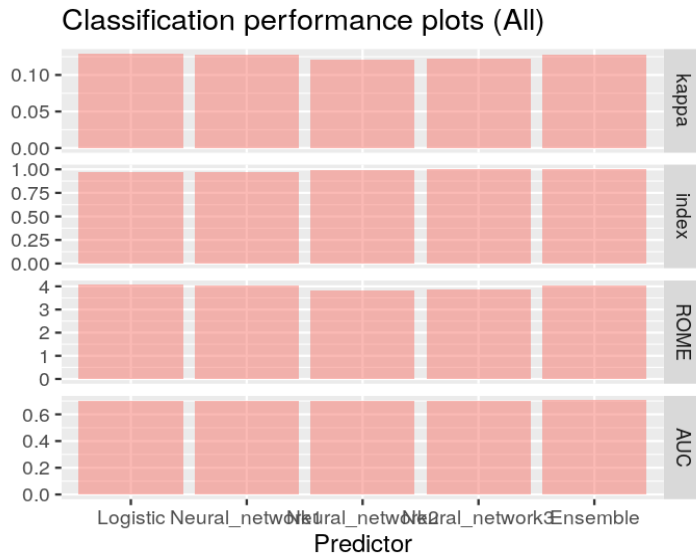
Cost:Margin: 1.41 : 60

Type	Predictor	TP	FP	TN	FN	total	TPR	TNR	precision	Fscore
All	Logistic	712	5,268	16,129	391	22,500	0.646	0.754	0.119	0.201
All	Neural_network1	718	5,365	16,032	385	22,500	0.651	0.749	0.118	0.200
All	Neural_network2	740	5,774	15,623	363	22,500	0.671	0.730	0.114	0.194
All	Neural_network3	741	5,733	15,664	362	22,500	0.672	0.732	0.114	0.196
All	Ensemble	737	5,514	15,883	366	22,500	0.668	0.742	0.118	0.200

Type	Predictor	accuracy	kappa	profit	index	ROME	contact	AUC
------	-----------	----------	-------	--------	-------	------	---------	-----

All	Logistic	0.748	0.129	34,288	0.968	4.067	0.266	0.700
All	Neural_network1	0.744	0.127	34,503	0.974	4.023	0.270	0.700
All	Neural_network2	0.727	0.121	35,215	0.995	3.834	0.290	0.701
All	Neural_network3	0.729	0.122	35,332	0.998	3.871	0.288	0.702
All	Ensemble	0.739	0.128	35,406	1.000	4.017	0.278	0.705

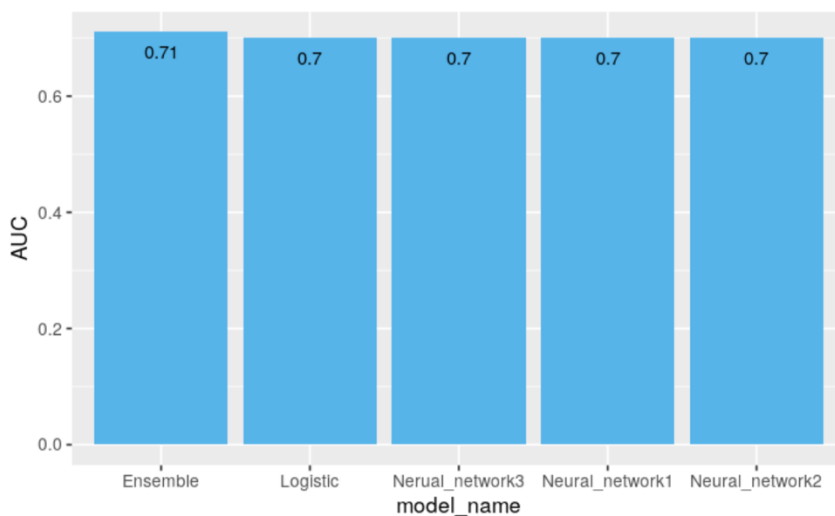
```
plot(compare_result, custom = FALSE)
```



6. Result Visualization

Visualize AUC Comparison

```
visualize(
  compare_result$dataset,
  xvar = "Predictor",
  yvar = "AUC",
  type = "bar",
  labs = list(title = "AUC", x = ""),
  custom = TRUE
) +
  geom_text(aes(label = format_nr(AUC, dec = 2)), vjust = 2)
```

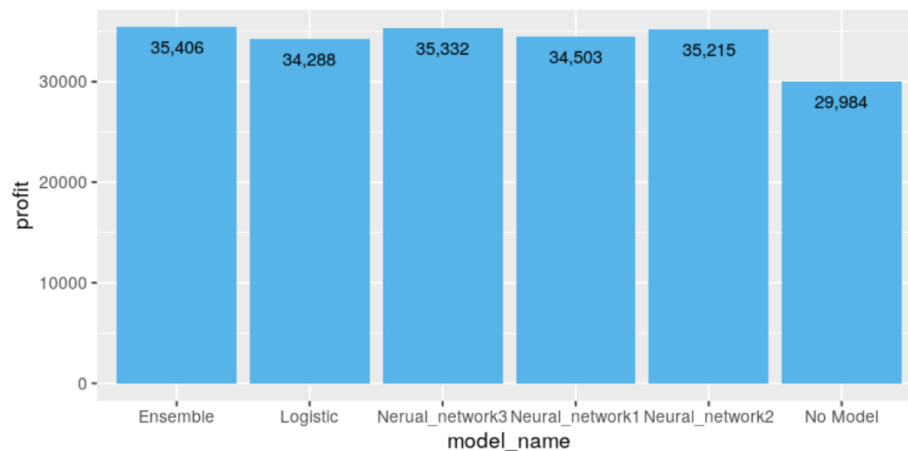


Visualize Profit Comparison

Benchmark Profit

```
benchmark_profit <- sum(testing$res1 == "Yes")*margin_sales - 22500 * mail_cost

visualize(
  compare_result$dataset,
  xvar = "Predictor",
  yvar = "profit",
  type = "bar",
  labs = list(title = "Profit", x = ""),
  custom = TRUE
) +
  geom_text(aes(label = format_nr(profit, dec = 0)), vjust = 2)
```



Conclusion:

1. Increased the R-square value of Logistic Regression by adding two new variables **zip_801** and **zip_804**. When starting out, I performed basic exploratory analysis to observe correlations and associations contained in the data. As seen above, I noted that the response rate of two zip codes, 00801 and 00804, in Wave 1 was high relative to the other zip codes. Therefore, I decided to target these two zip codes in Wave 2, and create two variables zip_801 and zip_804.
2. Increased the R-square value of Logistic Regression by adding two **interaction effects**. I attempted to find interactions between the variables through the performance of the Neural Network. And I found that there are interaction effects between version1 and last and numords.
3. Compared and evaluated different models based upon the profit and Area-Under-the-Curve (AUC). These results can be seen above, the **ensemble** model performed the best in terms of both profit and AUC value.
4. Increased upsell campaign profits by **18%** $(35406 - 29984)/35406$ by using the created ensemble model to develop email campaign strategies