**Homework 5 Assignment**
**By Group 9**

# Homework 5
# Group 9

**Members**

Rahil Shah, Parveen Kumar, Jainam Gandhi, Shwet Shah, Bhavya Shah, Nihar Khillar

# Homework 5 Assignment
# By Group 9

**When you have many variables how do you short list a few important variables that could affect churn?**

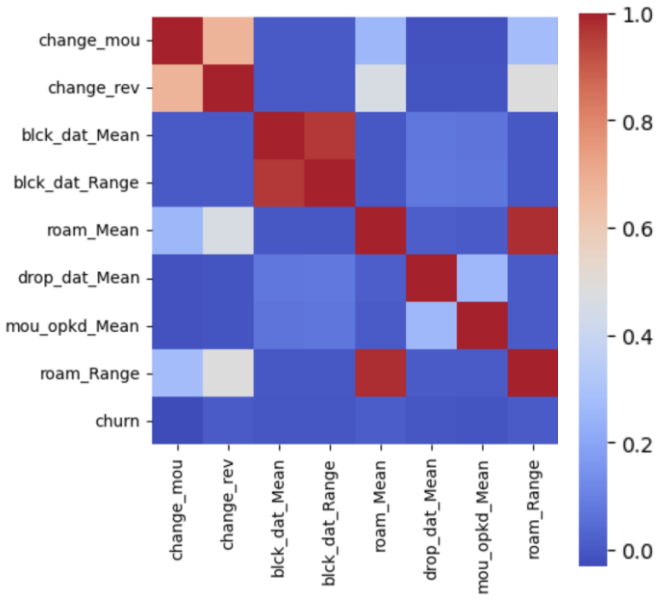
1. **Compute means of all x variables for churners and non-churners separately and compute the percentage difference for each variable.**
2. **Sort the variables from high to low based on percentage difference in means. The top 10 variables based on this criterion are good candidates for inclusion in the logit model if they are not highly correlated. When sorting variables note that the largest percentage differences are found both at the top and at the bottom.**
3. **Make sure that your data is not reduced by a large percentage (80-90%) because you have included explanatory variables in your model that have many missing values. Managers expect you will use most of the data (that is 80-90% of the data).**
4. **Create a random sample of 70,000 customers using PROC SURVEYSELECT in SAS. Call this the estimation sample (training sample).**
5. **Create a holdout sample with the rest of the 30000 customers (test sample).**
6. **Use a logistic regression model to build the best model (that is best in terms of model fit criteria AIC, BIC). Make sure that no two explanatory variables are highly correlated. Use correlation analysis to determine the correlation between the variables.**

Data preparation steps:

We did the data pre-processing in Python. The following steps were followed:

a. Means of churners and non-churners were calculated for all numerical variables
b. The top 10 numerical variables with highest absolute difference between the means of churners and non-churners were identified
c. Retdays and rmev were found to have missing values in excess of 10% and would be excluded from further analysis
d. Correlation was obtained for all pairs. For pairs with correlation over 0.7, one of the variables from such pairs would be omitted from further analysis (roam_mean and roam_range & blck_dat_mean and blck_dat_range had correlation over 0.9)

# Homework 5 Assignment
# By Group 9




e. The final list of numerical variables which would be used in logistic regression later were: change_mou, change_rev, blck_dat_Mean, roam_Mean, drop_dat_Mean, mou_opkd_Mean

f. Using Chi-sq test, we found the top 10 categorical variables which have a strong relationship with churn

g. Categorical variables (last_swap and hnd_webcap) which had missing values in excess of 10% were identified and would be excluded from further analysis

h. Chi-sq test was done for all the possible pairs of the balance 8 categorical variables. Area and csa were found to have a significant relation and we decided to drop area from further analysis

i. We trimmed the dataset by keeping the only the columns which were found relevant which have significant relation with churn. The final list of numerical and categorical variables which would be used for further analysis is:

```
change_mou
change_rev
blck_dat_Mean
roam_Mean
drop_dat_Mean
mou_opkd_Mean
csa
crclscod
asl_flag
ethnic
dualband
refurb_new
marital
```

# Homework 5 Assignment
# By Group 9

All the further steps were performed in SAS

a. The modified dataset containing only the relevant variables identified above was imported in SAS
b. The total observations came down to 97,351 from the original 100,000 since the missing values had been omitted
c. Using SURVEYSELECT, the dataset was split in the ratio of 70:30 i.e. 70% data in the training dataset and 30% data in the testing dataset

| Frequency<br>Percent<br>Row Pct<br>Col Pct | Table of Selected by churn | | |
|---|---|---|---|
| Selected(Selection Indicator) | churn | | |
| | 0 | 1 | Total |
| 0 | 14793<br>15.20<br>50.65<br>30.00 | 14412<br>14.80<br>49.35<br>30.00 | 29205<br>30.00 |
| 1 | 34518<br>35.46<br>50.65<br>70.00 | 33628<br>34.54<br>49.35<br>70.00 | 68146<br>70.00 |
| Total | 49311<br>50.65 | 48040<br>49.35 | 97351<br>100.00 |

d. Using PROC LOGISTIC, several models were developed and the best model was obtained based on higher likelihood ratio. This model would be used for answering the homework question.

# Homework 5 Assignment
# By Group 9

1. **Include a clean table of coefficients, t-values, and odds ratio only. I do not want the entire SAS output. Interpret the logistic output explaining AIC/BIC, meaning of coefficients, significance of betas, prediction accuracy (percent concordance), odds-ratios etc.**

| Model Information | |
|---|---|
| Data Set | WORK.CHURCH_TRAINING_V1 |
| Response Variable | churn |
| Number of Response Levels | 2 |
| Model | binary logit |
| Optimization Technique | Fisher's scoring |

| | |
|---|---|
| Number of Observations Read | 68146 |
| Number of Observations Used | 68146 |

| Response Profile | | |
|---|---|---|
| Ordered Value | churn | Total Frequency |
| 1 | 1 | 33628 |
| 2 | 0 | 34518 |

Probability modeled is churn='1'.

The training dataset had 68,146 observations out of which 33,648 rows had churn value 1 (customer left between 31-60 days after the obs_date) while 34,518 rows had churn value 0.

| Model Fit Statistics | | |
|---|---|---|
| Criterion | Intercept Only | Intercept and Covariates |
| AIC | 94460.792 | 93623.717 |
| SC | 94469.921 | 93742.399 |
| -2 Log L | 94458.792 | 93597.717 |

| Testing Global Null Hypothesis: BETA=0 | | | |
|---|---|---|---|
| Test | Chi-Square | DF | Pr > ChiSq |
| Likelihood Ratio | 861.0748 | 12 | <.0001 |
| Score | 830.1490 | 12 | <.0001 |
| Wald | 830.9077 | 12 | <.0001 |

AIC and BIC: Akaike's Information Criterion (AIC) and Schwartz's Criterion (SC) are model fit measures. AIC is equal to -2Log L+2p and SC is equal to -2Log L+p*log (n) where p is the number

# Homework 5 Assignment
# By Group 9

of parameters and n is the number of observations in the dataset. AIC penalizes for the number of parameters while SC is a stricter measure since it penalizes for both number of parameters and number of observations. Models with lower AIC and SC are better. Our model's AIC and SC are lower for the intercepts and covariates model compared to the null model indicating that the model improves with the addition of the selected covariates.

The likelihood ratio is 861 (difference in -2logL of null model & Intercept and covariates model). The p value for the Chi-sq test is less than 0.001. Considering a confidence interval of 95% or critical alpha of 0.05 for the test, we reject the null hypothesis (no improvement over the null model) and conclude that our model significantly improves over the null model.

The McFadden's R-sq= Diff. in -2Log L/ -2Log L of Null model= 861.07/94458.79= 0.9% (Our model has a 0.9% improvement over the null model)

**Type 3 Analysis of Effects**

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|
| change_mou | 1 | 168.2254 | <.0001 |
| change_rev | 1 | 85.4067 | <.0001 |
| roam_Mean | 1 | 15.3826 | <.0001 |
| asl_flag | 1 | 369.8674 | <.0001 |
| dualband | 3 | 160.8319 | <.0001 |
| refurb_new | 1 | 61.2245 | <.0001 |
| marital | 4 | 85.3896 | <.0001 |

The Type 3 Analysis of effects is like Type 3 sum of squares in GLM. The p values of all the variables are less than 0.0001 i.e. all the variables are significant at 95% confident interval.

# Homework 5 Assignment
# By Group 9

| Parameter | Class helper | Estimate | P value | Odds ratio |
|---|---|---|---|---|
| Intercept | | -0.22 | <0.0001 | 0.999 |
| change_mou | | -0.0005 | <0.0001 | 1.002 |
| change_rev | | 0.0025 | <0.0001 | 1.004 |
| roam_Mean | | 0.0043 | <0.0001 | 1.555 |
| asl_flag | N vs Y | 0.44 | <0.0001 | 1.627 |
| dualband | N vs Y | 0.19 | <0.0001 | 1.255 |
| dualband | T vs Y | -0.22 | <0.0001 | 0.806 |
| dualband | U vs Y | -0.43 | 0.0096 | 0.650 |
| refurb_new | N vs R | -0.17 | <0.0001 | 0.842 |
| marital | A vs U | -0.10 | 0.0043 | 0.903 |
| marital | B vs U | -0.07 | 0.0218 | 0.930 |
| marital | M vs U | -0.15 | <0.0001 | 0.859 |
| marital | S vs U | -0.16 | <0.0001 | 0.853 |

All the variables have p value less than critical p value of 0.05, hence all the variables are significant (null hypothesis of no significance is rejected)

Interpretation of the coefficients/ odds ratio in the above table:

1. Change_rev: For a unit increase in change_rev, the log odds of churn (compared to no churn) increases by 0.0025. For a unit increase in change_rev, the odds of churn increases by 0.4%

Interpretating the odds ratio makes more sense than interpretating the coefficients since coefficients denote the change in log odds which is more difficult to interpret in business context

2. Change_mou: For a unit increase in change_mou, the odds of churn increases by 0.2%
3. Roam_Mean: For a unit increase in roam_mean, the odds of churn increases by 55.5%
4. Asl_flag N vs Y: When account spending limit is N, the odds of churn increases by 62.7% compared to when it is Y
5. Dualband N vs Y: When dualband is N, the odds of churn increases by 25.5% compared to when it is Y
6. Refurb_new: When handset is new, the odds of churn decreases by 15.8% compared to when it is refurbished
7. Marital A vs U: When marital status is A, the odds of churn decreases by 9.7% compared to when it is U

# Homework 5 Assignment
# By Group 9

| Association of Predicted Probabilities and Observed Responses | | | |
|---|---|---|---|
| Percent Concordant | 56.7 | Somers' D | 0.135 |
| Percent Discordant | 43.3 | Gamma | 0.135 |
| Percent Tied | 0.0 | Tau-a | 0.067 |
| Pairs | 1160771304 | c | 0.567 |

Percent concordant: Of the 1.16+ billion pairs possible from 2 groups (churn=1 and churn=0), 56.7% pairs have probability of churn=1 (event) higher than the probability of churn=0 (non event). Higher the concordance, better the model. In our model, the obtained concordant % of 56.7% which is higher than 50.6% which is the naïve guess probability.

2. **Which are the top three factors that affect churn in your model and what is their effect size?**

| Analysis of Maximum Likelihood Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| Parameter | | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq | Standardized Estimate |
| Intercept | | 1 | -0.2172 | 0.0297 | 53.4676 | <.0001 | |
| change_mou | | 1 | -0.00051 | 0.000039 | 168.2254 | <.0001 | -0.0799 |
| change_rev | | 1 | 0.00249 | 0.000270 | 85.4067 | <.0001 | 0.0747 |
| roam_Mean | | 1 | 0.00432 | 0.00110 | 15.3826 | <.0001 | 0.0394 |
| asl_flag | N | 1 | 0.4417 | 0.0230 | 369.8674 | <.0001 | 0.0841 |
| dualband | N | 1 | 0.1913 | 0.0185 | 107.2110 | <.0001 | 0.0445 |
| dualband | T | 1 | -0.2162 | 0.0389 | 30.9385 | <.0001 | -0.0242 |
| dualband | U | 1 | -0.4311 | 0.1665 | 6.6994 | 0.0096 | -0.0114 |
| refurb_new | N | 1 | -0.1715 | 0.0219 | 61.2245 | <.0001 | -0.0333 |
| marital | A | 1 | -0.1022 | 0.0358 | 8.1660 | 0.0043 | -0.0126 |
| marital | B | 1 | -0.0724 | 0.0315 | 5.2623 | 0.0218 | -0.0102 |
| marital | M | 1 | -0.1522 | 0.0187 | 65.9347 | <.0001 | -0.0391 |
| marital | S | 1 | -0.1589 | 0.0222 | 51.3702 | <.0001 | -0.0336 |

Observing the standardized estimates, the top 3 factors that affect churn in the model are asl_flag, change_mou and change_rev. Looking at the odds ratio in the table given in Ques 1, we can interpret:

1. Asl_flag N vs Y: When account spending limit is N, the odds of churn increases by 62.7% compared to when it is Y
2. Change_mou: For a unit increase in change_mou, the odds of churn increases by 0.2%
3. Change_rev: For a unit increase in change_rev, the odds of churn increases by 0.4%

3. **What other variables (that if collected) would help to improve the fit of the model.**

# Homework 5 Assignment
# By Group 9

Other variables which might improve the fit of the model are as under:

a. Prepaid/ Postpaid line: The exit barrier is higher in a postpaid line compared to a prepaid line, hence, prepaid consumers might have higher churn rate compared to postpaid consumers.
b. Number of household lines: If a single household has subscribed to family plan or has taken multiple lines under a single account, then the churn rate for such customers should be lower compared to single line customers
c. Whether customer uses other bundled services like wifi from the same company: Customers who use multiple services from the same provider are expected to have lower churn rate compared to the ones who just subscribe to the phone line
d. Net promoter score (NPS): NPS is a customer loyalty metric which is generally taken by asking the customers how likely they are to recommend the telecom's products and services to others on a scale of 1-10. Customers with higher NPS are expected to have lower churn rate.
e. Engagement in the telecom apps: If a customer spends more time in the telecom application, then such customers have higher engagement and are expected to have lower churn rate. It is important to note the sections of the application where a customer spends more time, for example, if a customer spends more time in the complaints section, then he/ she is likely to be unhappy with the services and hence more likely to quit even though engagement may be high.

4. **Compute the hit ratio for your model. Hit ratio is defined as the percentage of correct predictions using the logit model. Use the model to predict 1 or 0 using the same data.**

The FREQ Procedure

| Frequency | Table of churn by pred_churn | | |
|---|---|---|---|
| | pred_churn | | |
| churn | 0 | 1 | Total |
| 0 | 19349 | 15169 | 34518 |
| 1 | 15664 | 17964 | 33628 |
| Total | 35013 | 33133 | 68146 |

# Homework 5 Assignment
# By Group 9

Hit ratio is % of events correctly classified. As per the above table:

Number of non churners correctly classified: 19,349 (1)

Number of churners correctly classified: 17,964 (2)

Total number of observations: 68,146 (3)

Hit ratio= ( (1) + (2) )/ (3)= (19,349 + 17,964)/ 68,146= 54.75%

5. **Using the model parameters predict the churn for the holdout sample as well and compute the hit ratio.**

The FREQ Procedure

| Frequency | Table of churn by pred_churn | | |
|---|---|---|---|
| | pred_churn | | |
| churn | 0 | 1 | Total |
| 0 | 8348 | 6445 | 14793 |
| 1 | 6542 | 7870 | 14412 |
| Total | 14890 | 14315 | 29205 |

Hit ratio is % of events correctly classified. As per the above table:

Number of non churners correctly classified: 8,348 (1)

Number of churners correctly classified: 7,870 (2)

Total number of observations: 29,205 (3)

Hit ratio= ( (1) + (2) )/ (3)= (8,348 + 7,870)/ 29,205 = 55.5%