Capstone Project Guide

Telecom Churn Data Set

Agenda

- Understanding Data: Creating a data quality report
- Variable Profiling: Continuous Variables
- Variable Profiling: Categorical Variables
- Data preparation: Binning, missing value imputation, derived variables and dummy variable creation
- Model Building: Using step wise to select variable
- Final model
- Creating customer segments
- Additional notes

Understanding data: Creating data quality report

- A data quality report with the format given below, needs to be created. Refer to the class recordings of “Data preparation and exploration” for approach and R codes
- F:\Work\Jigsaw Academy\Data Scientist Course\Data Science Redo\Telecom Case study\Sample Data quality report.xlsx
- Variables with a lot of missing values will be omitted from the analysis

Variable Profiling: Continuous Variables

- Decile binning should be used to find out event rate (churn rate in this case). If there is a trend (increase in event rate or decrease in event rate), then that variable should be selected for model iterations
- For some variables it won't be possible to do decile binning, one should either divide the data into 8,6,4,3 or 2 equal parts and then look at the event rates. If there is a pattern then, the variable should be chosen for model iterations
- Profiling will also aid in data preparation as deciles with similar event rates can be clubbed together into one group and a categorical variable can be created

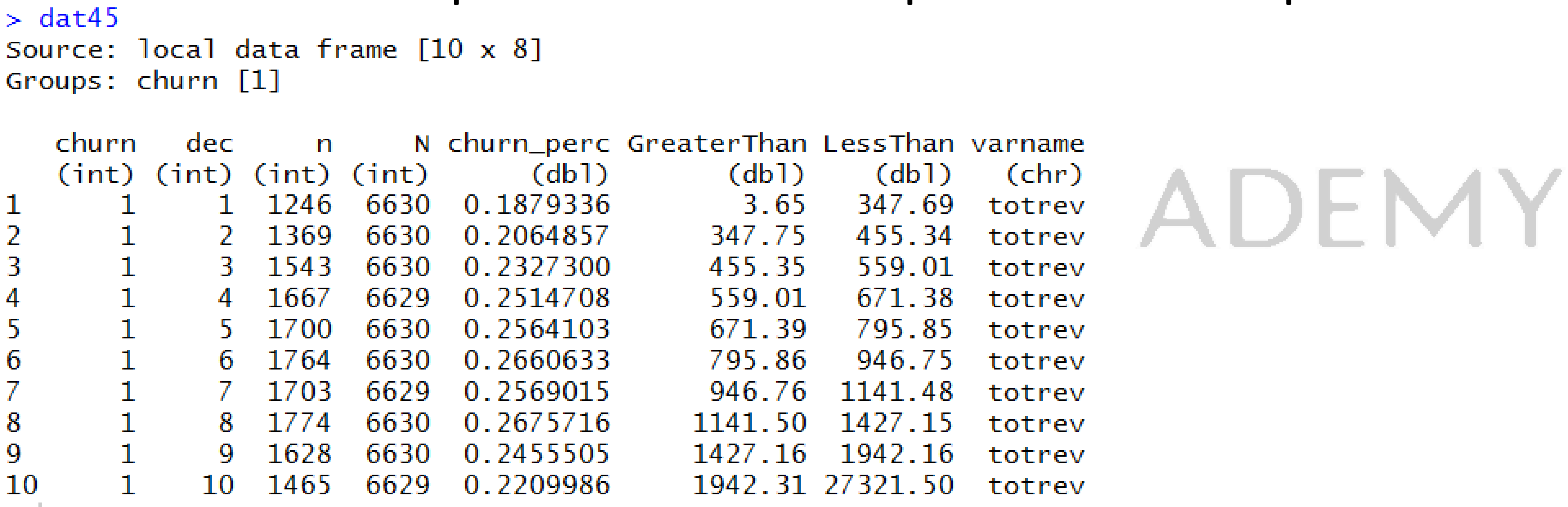
Variable Profiling: Continuous Variables

- Following is an R code sample that can be used to do decile binning. One can refer to pre class video in “Data preparation” for more details on the ntile() function used below:

```
telecom%>%mutate(dec=ntile(totrev,n=10))%>%count(churn,dec)%>%filter(churn==1)->dat45

dat45$N<-unclass(telecom%>%mutate(dec=ntile(totrev,n=10))%>%count(dec)%>%unname())[[2]]

dat45$churn_perc<-dat45$n/dat45$N

dat45$GreaterThan<-unclass(telecom%>%mutate(dec=ntile(totrev,n=10))%>%group_by(dec)%>%summarise(min(totrev)))[[2]]

dat45$LessThan<-unclass(telecom%>%mutate(dec=ntile(totrev,n=10))%>%group_by(dec)%>%summarise(max(totrev)))[[2]]

dat45$varname<-rep("totrev",nrow(dat45))
```

Variable Profiling: Continuous Variables

- The code in the previous slide will produce an output like this:



```
> dat45
Source: local data frame [10 x 8]
Groups: churn [1]

   churn   dec     n     N churn_perc GreaterThan LessThan varname
   (int) (int) (int) (int)      (dbl)       (dbl)    (dbl)   (chr)
1      1     1  1246  6630  0.1879336        3.65   347.69  totrev
2      1     2  1369  6630  0.2064857      347.75   455.34  totrev
3      1     3  1543  6630  0.2327300      455.35   559.01  totrev
4      1     4  1667  6629  0.2514708      559.01   671.38  totrev
5      1     5  1700  6630  0.2564103      671.39   795.85  totrev
6      1     6  1764  6630  0.2660633      795.86   946.75  totrev
7      1     7  1703  6629  0.2569015      946.76  1141.48  totrev
8      1     8  1774  6630  0.2675716     1141.50  1427.15  totrev
9      1     9  1628  6630  0.2455505     1427.16  1942.16  totrev
10     1    10  1465  6629  0.2209986     1942.31 27321.50  totrev
```

- All such objects containing profiles of the continuous variables can be written out as a csv file for further analysis

Variable Profiling: Categorical Variables

- Event rate for each level in a categorical variable can be computed
- Ideally there should be good difference between the event rate in each level
- If some levels have similar event rate then those labels can be combined in a single group

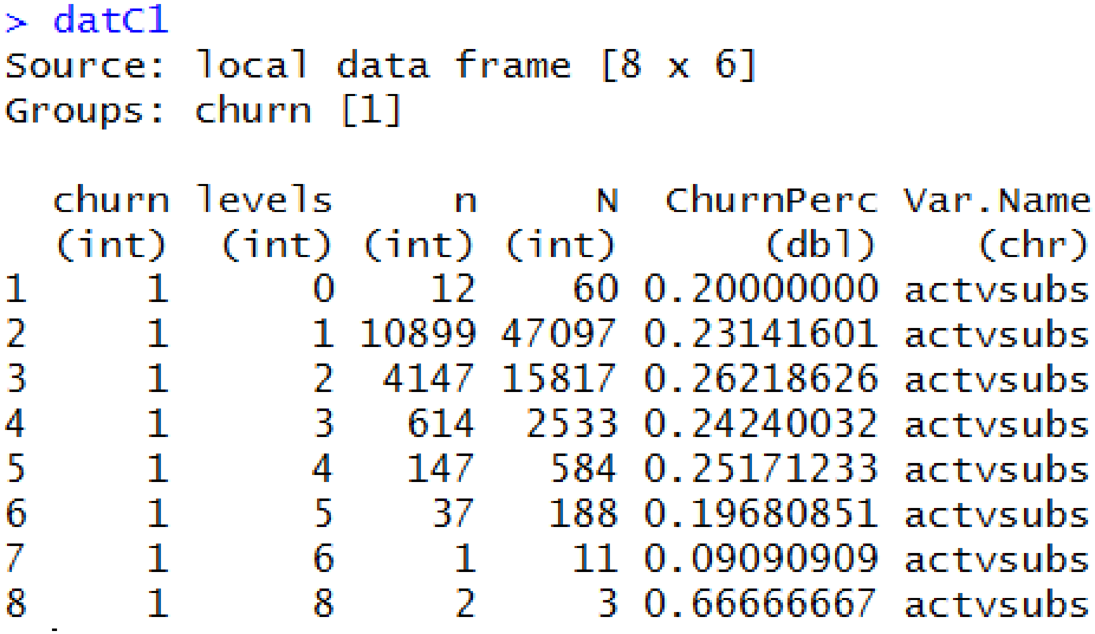
Variable Profiling: Categorical Variables

- This is a sample R code that can be used to create categorical variable profiles:

```
telecom%>%count(churn,levels=actvsubs)%>%filter(churn==1)->datC1

datC1$N<-unclass(telecom%>%filter(actvsubs%in%datC1$levels)%>%count(actvsubs))[[2]]

datC1$ChurnPerc<-datC1$n/datC1$N

datC1$Var.Name<-rep("actvsubs",nrow(datC1))
```

Variable Profiling: Categorical Variables

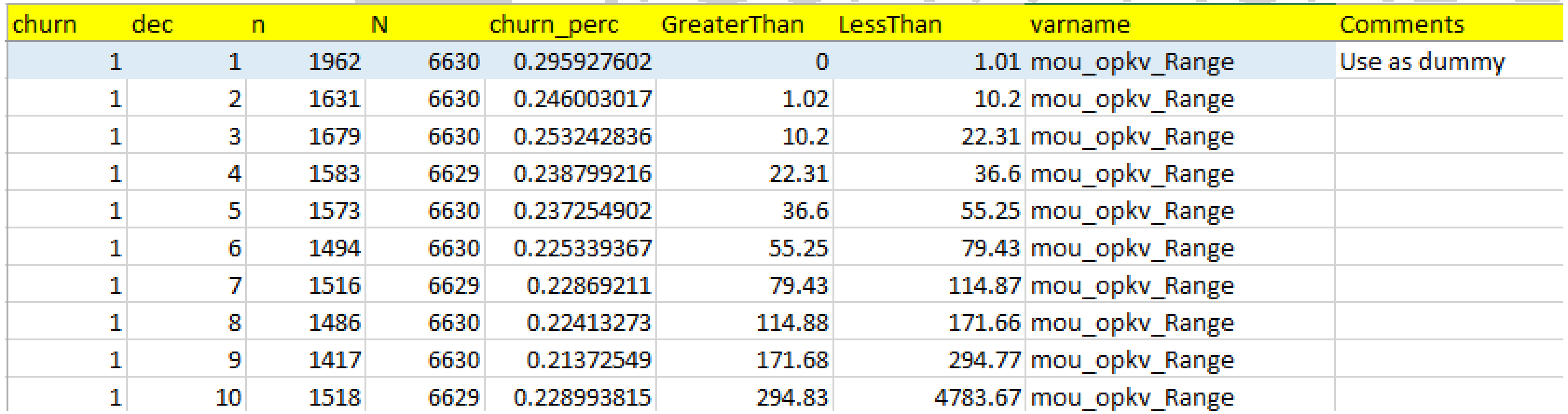
- An output like this will be produced using the code in the previous slide:



```
> datC1
Source: local data frame [8 x 6]
Groups: churn [1]

  churn levels     n     N  ChurnPerc Var.Name
  (int)  (int) (int) (int)      (dbl)    (chr)
1     1      0    12    60 0.20000000 actvsubs
2     1      1 10899 47097 0.23141601 actvsubs
3     1      2  4147 15817 0.26218626 actvsubs
4     1      3   614  2533 0.24240032 actvsubs
5     1      4   147   584 0.25171233 actvsubs
6     1      5    37   188 0.19680851 actvsubs
7     1      6     1    11 0.09090909 actvsubs
8     1      8     2     3 0.66666667 actvsubs
```

Data Preparation: Continuous Variables

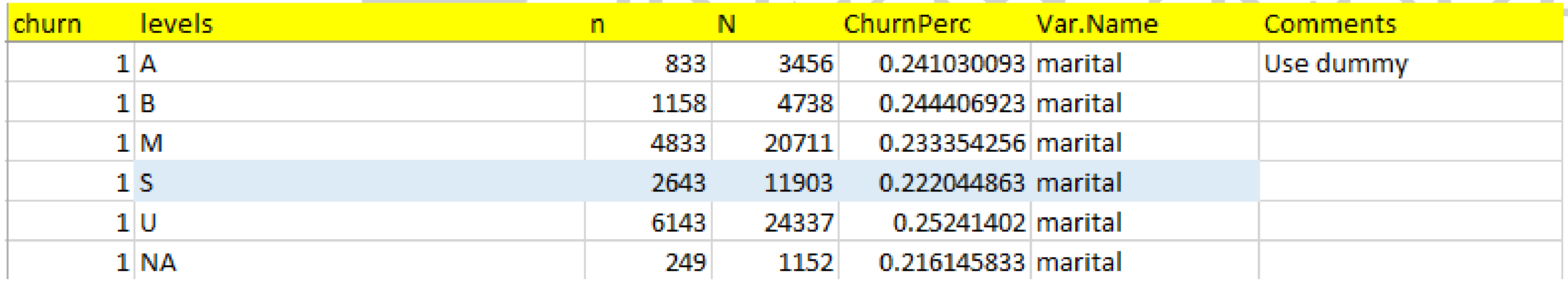
- Based on the variable profiles created in the previous steps, data preparation can be done.
- Some continuous variables can be converted to dummy variables based on the similarity of event rates.



churn	dec	n	N	churn_perc	GreaterThan	LessThan	varname	Comments
1	1	1962	6630	0.295927602	0	1.01	mou_opkv_Range	Use as dummy
1	2	1631	6630	0.246003017	1.02	10.2	mou_opkv_Range	
1	3	1679	6630	0.253242836	10.2	22.31	mou_opkv_Range	
1	4	1583	6629	0.238799216	22.31	36.6	mou_opkv_Range	
1	5	1573	6630	0.237254902	36.6	55.25	mou_opkv_Range	
1	6	1494	6630	0.225339367	55.25	79.43	mou_opkv_Range	
1	7	1516	6629	0.22869211	79.43	114.87	mou_opkv_Range	
1	8	1486	6630	0.22413273	114.88	171.66	mou_opkv_Range	
1	9	1417	6630	0.21372549	171.68	294.77	mou_opkv_Range	
1	10	1518	6629	0.228993815	294.83	4783.67	mou_opkv_Range	

Data Preparation: Categorical Variables

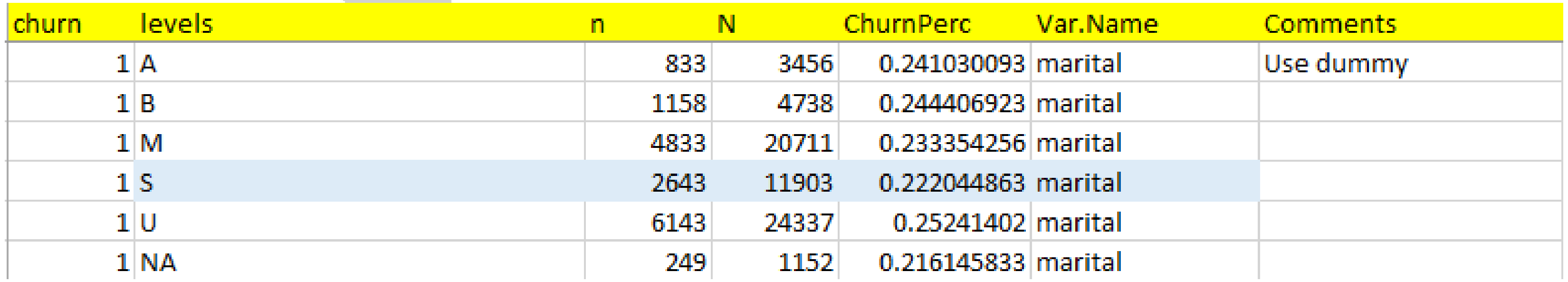
- There are categorical variables with several levels, based on the similarity in event rate, these levels can be reduced.
- Ideally one should not use a variable with more than 3 levels, if a variable has a lot of levels then they should be reduced.



churn	levels	n	N	ChurnPerc	Var.Name	Comments
1	A	833	3456	0.241030093	marital	Use dummy
1	B	1158	4738	0.244406923	marital	
1	M	4833	20711	0.233354256	marital	
1	S	2643	11903	0.222044863	marital	
1	U	6143	24337	0.25241402	marital	
1	NA	249	1152	0.216145833	marital	

Data Preparation: Missing Value Imputation

- There are several variables with missing values, based on the percentage of missing values in these variables, some variables can be out rightly excluded from analysis
- For the variables with a few missing values, imputation can be made by observing the event rate



churn	levels	n	N	ChurnPerc	Var.Name	Comments
1	A	833	3456	0.241030093	marital	Use dummy
1	B	1158	4738	0.244406923	marital	
1	M	4833	20711	0.233354256	marital	
1	S	2643	11903	0.222044863	marital	
1	U	6143	24337	0.25241402	marital	
1	NA	249	1152	0.216145833	marital	

- Here missing values can be imputed as “S”

Data Preparation: Derived Variables

- To answer some of the issues raised in the case, additional variables will need to be created for example to understand if “Network issues are leading to churn” following variable can be created

Completion_Percentage=Completed Voice Calls/Total Placed Calls

- There are several such issues which have been raised in the case study for which new variables will have to be created.

Model Building: Using stepwise regression

- Split the data into test and training sets
- Make sure this split is random
- If results are to be reproducible then set.seed() can be used
- One can use step() function to do a stepwise regression and choose variables
- Even after doing a stepwise regression, one would need to run several iterations before the model is finalised

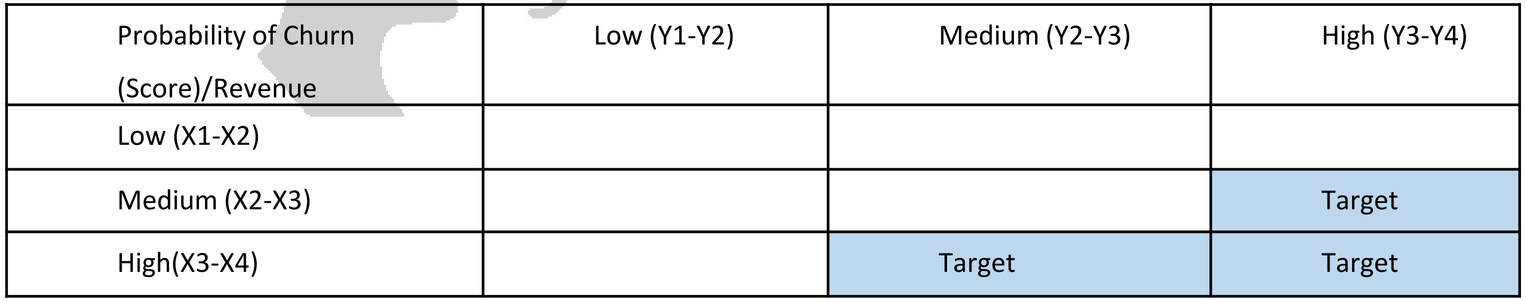
Final Model

- The Final model should:

1. Include variables with significant beta coefficients
2. Have variables where beta coefficients have a proper sign
3. Have a good fit measured by AUC or Concordance (For this data an AUC of 0.63-0.62 is attainable)

Creating customer segments

- Once the model has been finalized, one of the tasks would be to create customer segments for proactive targeting.
- Using the logistic model, one can easily compute the probability of a customer churning and come up with a grid like this:



Probability of Churn (Score)/Revenue	Low (Y1-Y2)	Medium (Y2-Y3)	High (Y3-Y4)
Low (X1-X2)			
Medium (X2-X3)			Target
High(X3-X4)		Target	Target

Additional Notes

- Make sure you answer the top line questions raised in the case study
- A lot of time will have to be spent on exploration/profiling and preparation of data
- Before attempting the final case study make sure you have completed all the previous case studies and quiz
- Budget sufficient time for submitting the case study and seeking query, last minute submissions and queries can't be prioritised and will lead to delays.
- Make sure final case study submission happens at least 20 days prior to your course expiry.