B.Sc in Computing

# Enterprise Database Technologies CA 1

Released: 24th February 2017

Upload by 24th March 2017

## Section 1 - Data Understanding and Data Exploration

### See R sheet (pre-processing question)

### See R sheet (pre-processing question)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Predictor | A | B | C | D | E | F | G |
| AREA\_CODE | Nominal | 0 | X | X | X | X | X |
| CUST\_MOS | Numeric | 3/2071 | **Min**: 1,  **Median**: 11,  **Mean**: 16.05,  **Max**:  50 | Most customers seem to stay during 5 to 15 months | It seems that in the first months, the customer has more chances to Churn, Around 10 months the customer will churn as well (end of one year contract?) | Skewness: 1.131244  Positively skewed (skewed to the right) | One outlier found in the box plot |
| LONGDIST\_FLAG | Nominal | 0 | X | X | X | X | X |
| CALLWAITING\_FLAG | Nominal | 0 | X | X | X | X | X |
| NUM\_LINES | Numeric | 0 | **Min**: 1,  **Median**: 1,  **Mean**: 1.391,  **Max**:  3 | Most users seem to have 1 number only and only very few have 3 numbers | The number of lines do not seem to bring much insight | Skewness: 2.057503  Strongly positively skewed (skewed to the right) | X |
| VOICEMAIL\_FLAG | Nominal | 0 | X | X | X | X | X |
| MOBILE\_PLAN | Nominal | 0 | X | X | X | X | X |
| CONVERGENT\_BILLING | Nominal | 0 | **Mode**: No | X | X | X | X |
| GENDER | Nominal | 0 | **Mode**:  M | X | Does not bring insight | X | X |
| INCOME | Ordinal | 0 | **Mode**:  Medium Income | Most users have medium income, twice as many users have high income compared to low income users | Users with low income or high income tend not to churn while medium incomes tend to churn more | X | X |
| PHONE\_PLAN | Ordinal | 4/2071 | Mode: **International** | Only few users choose the Euro-zone, most of the users opt for the International and National plans | Users having the Euro-Zone or the International phone plan tend to churn while users with a National or Promo\_plan tend to churn less | X | X |
| EDUCATION | Nominal | 8/2071 | Mode: **Post Primary** | Post-Primary is the dominant group, the number of High school and Primary school are very low | Primary are churners, Masters tend to churn, PhD tend not to churn | X | X |
| TOT\_MINUTES\_USAGE | Numeric | 4/2071 | **Min**: 0,  **Median**: 264,  **Mean**: 2036,  **Max**:  36237 | Clear majority of users use less than 2500 mins | Do not seem to bring insight | Skewness: 1.088757  Positively skewed (skewed to the right) | Graphically, from the boxplot we can see that the data contains a lot of outliers that will need to be cleaned out |

4.

TOTAL\_MINUTES\_USAGE:

IQR method: 176 outliers

Z standardisation method: 69 outliers

## Section 2 – Data Mining