Churn Rate Prediction

Analysis of Telecom Dataset

By Mano Bharathi J V

Agenda

- Project Synopsis
- Understanding Data
- Data Preparation: Missing value Imputation, Dummy Variable Creation
- Feature Engineering: Derived variables creation
- Model Building: Using step wise model
- Final Model: Iterative process of identifying Significant variables
- Top Line Questions of Interest
- Creating Customer Segments

Project Synopsis

- **Background:**

Telecom industry level survey reports have been just released. Mobicom is concerned that Market environment of rising churn and declining ARPU will hit them hard as churn rate of customers is relatively high at Mobicom. Currently they are using a reactive strategy when the customer calls in to close the account.

- **Objective:**

Mobicom is planning to roll out targeted pro-active retention programs. To be able to effectively drive these retention strategies, a few questions of interest requires data-based insights and recommendations relating to customer churn.

- **Available Data:**

Data of mature customers who were with the telecom for atleast six months were sampled and predictor variables were calculated based on the previous four months. Churn was calculated based on whether the customer left the company during 31-60 days after they were originally sampled.

Understanding Data

- Created a Data quality report for all the variables with columns: variable name, Datatype, NoOfRecords, UniqueRecords, DataAvailablePercent, MissingPercent and summary details like Mean, Median etc..
- 14 predictor variables have more than 10% missing values.
- Except 3, the remaining 11 variables were ignored since we couldn’t attribute any specific meaning for missing values. The missing values are simply unknown.
- Exception variables are as follows:
  - solflag - Missing values indicate default preference, Mostly the telecoms will setup this flag as 'N' by default.
  - retdays - Missing values for this variable means there have been no retention calls made by the customer in the past.
  - div_type - Represents additional services, Maybe the missing values represent No Additional services subscribed.

Data Preparation

Missing Value Imputation

- There are several variables with missing values.
- Missing values were imputed based on the churn rate association.
- Few categorical variables were ignored since there's no much difference in churn rate between the groups.
- Also categorical variable 'CSA' was ignored since there too many groups associated.
- Categories were grouped inside categorical variables when the churn rate was similar or close.
- Several continuous variables with missing values were imputed based on the median values since most of the continuous variables have shown right skewness.

Dummy Variable creation

- Created dummy variables using dummyVars() function for all categorical variables.
- Only n-1 dummy variables were created to avoid multi-collinearity issue.

Feature Engineering

- Created new variables out of existing predictor variables to help us aid in model accuracy and answer some to line business questions.
- Below are the variables and explanation:
  - fe_ovrmou_per – Percentage of overage minutes of use out of total minutes of use.
  - fe_vce_ovrrev_per – Percentage of voice overage revenue out of total overage revenue.
  - fe_vce_drop_per – Percenatge of voice calls dropped.
  - fe_dat_drop_per – Percenatge of data calls dropped.
  - fe_comp_drop_per – Overall drop percentage.
  - fe_diff36mou – Percentage difference in average minutes of use between 3 and 6 month's.
  - fe_diff3mou – Percentage difference in average minutes of use between 3 and entire life of customer.
  - fe_diff6mou – Percentage difference in average minutes of use between 6 and entire life of customer.

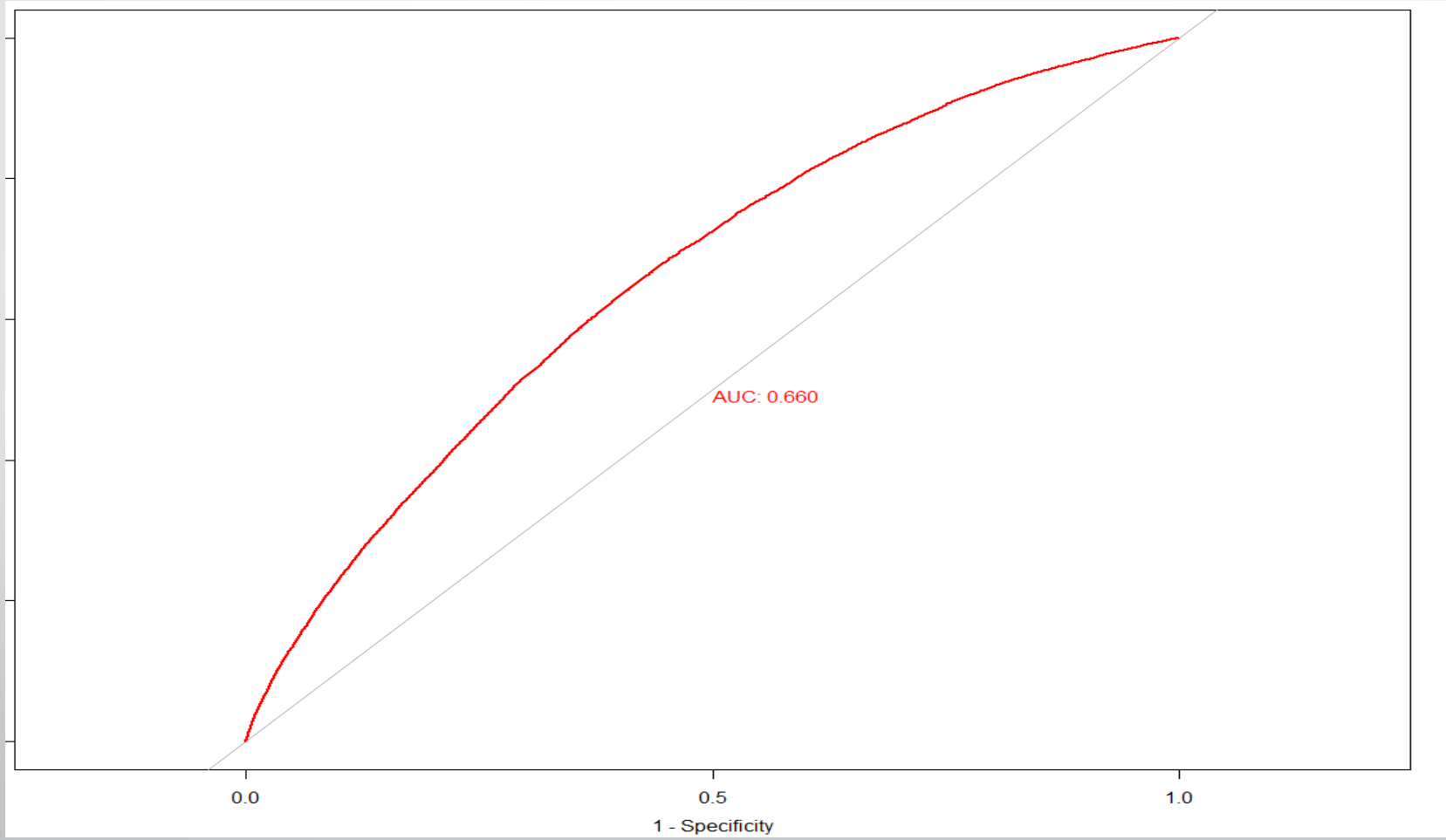
Model Building

- Converted continuous variables eqpdays and months into categories based on churn rate grouping.
- Divided the total data into 75% train and 25% test datasets.
- Using stepwise function, and direction as ‘both’, significant variables were identified.
- The variables were retrieved from the stepwise function using the formula parameter.

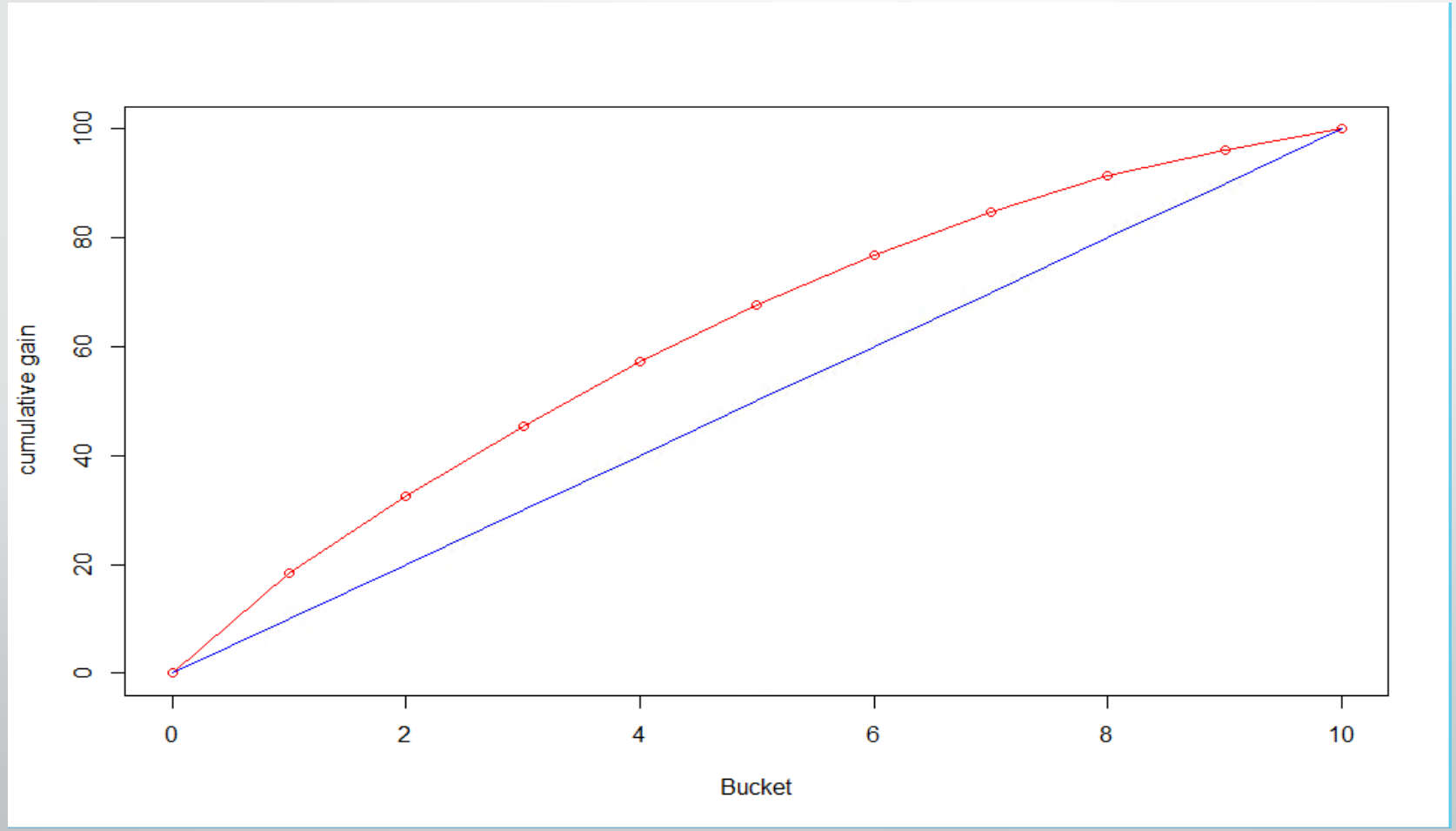
Final Model

- Even after using stepwise function, there were few insignificant predictor variables present in the model.
- Each were identified using the p-value and removed from the modelling iteratively.
- As a final step, model parameters were checked for VIF(Variable Inflation factor) and the predictor variables with VIF greater than 5 were removed iteratively.
- The significant variables were selected using the above steps and the final model yielded an AUC of 0.66 for the test dataset.
- Using the significant predictor variables finalised on the test dataset and cross fold validation with 4 folds, we predicted the churn probability of each customer in the entire dataset.

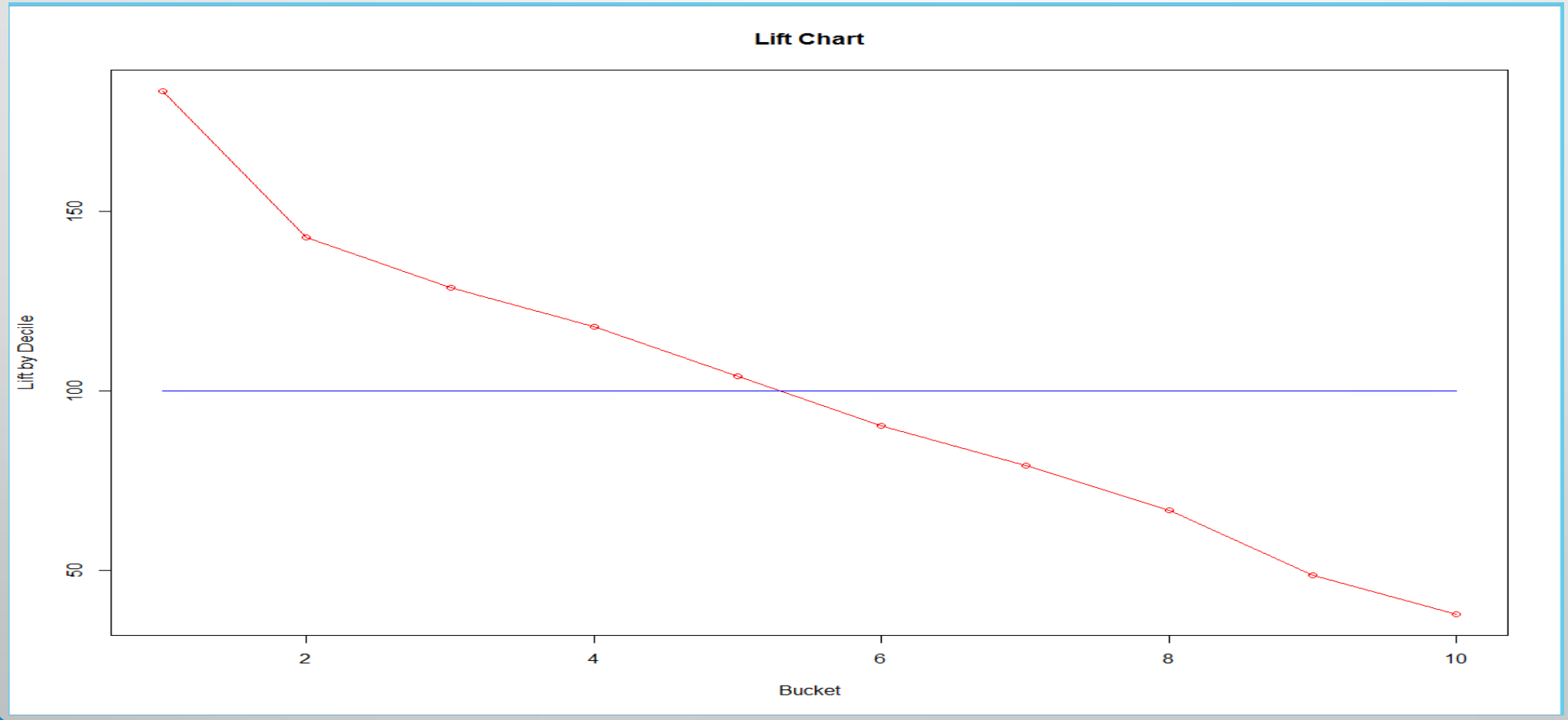
Final Model: AUC plot



Final Model: Gains Chart



Final Model: Lift Chart



Final Model: Gains table

A tibble: 10 x 9

	bucket	total	resp	resp_perc	max_prob	cumresp	cumGain	Lift	cumLift
	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	6612	2897	0.438	0.805	2897	18.3	183.	183.
2	2	6612	2255	0.341	0.366	5152	32.6	143.	163.
3	3	6611	2032	0.307	0.321	7184	45.5	129.	152.
4	4	6612	1863	0.282	0.291	9047	57.3	118.	143.
5	5	6611	1644	0.249	0.264	10691	67.7	104.	135.
6	6	6612	1426	0.216	0.238	12117	76.7	90.3	128.
7	7	6612	1252	0.189	0.212	13369	84.7	79.3	121.
8	8	6611	1054	0.159	0.183	14423	91.3	66.8	114.
9	9	6612	769	0.116	0.148	15192	96.2	48.7	107.
10	10	6611	598	0.0905	0.106	15790	100	37.9	100

- Top 20% of the probabilities lie between 0.321-0.805. It contains 32.6% of the churn customers.

Top Line Questions of Interest

- Top 5 factors driving churn:
  - retdays: Retention days- Number of days since the last retention call is made. It is the most important factor. When retdays changes from greater than 60 to less than 60, the log odds of churn increases by 0.194. When retdays changes from ‘NA’ to less than 60, the log odds of churn increases by 0.276.
  - months: Total number of months in service. Second most important factor. When months changes from (<=10 months and >33 months) to 11-13 months, the log odds of churn increases by 0.12. When months changes from (<=10 months and >33 months) to 14-33 months, the log odds of churn increases by 0.029.

Top Line Questions of Interest

- Top 5 factors driving churn:
  - eqpdays: Equipment days- Number of days of current equipment. It is the third important factor. When eqpdays changes from less than 300 days to greater than 300, the log odds of churn increases by 0.077.
  - uniqsubs: Number of unique subscribers in household. Fourth important factor. When uniqsubs changes from 1 to Others, the log odds of churn increases by 0.055.
  - asl_flag: Account spending limit restriction. Fifth important factor. If there's account spending limit restriction(flag='Y), the log odds of churn decreases by 0.047

Top Line Questions of Interest

- Validation of Survey Findings:
  - “cost and billing”, “network and service quality” are important factors?
    - “cost, billing and network quality” factors are not that important.
    - “service quality” is very important factor as depicted by retdays column earlier. The company should give extra attention to the customers who make retention calls to rectify their complaints.

Top Line Questions of Interest

- Validation of Survey Findings:
  - Are data usage connectivity issues turning out to be costly? Is it influencing churn?
    - Even though drop in data increasing the churn, but the variable is not much important and also just 15% of the entire customers are making use of network data. Since the global survey reports data usage as an important factor, the telecom should try to come up with strategic plans to make more customers using the data.
- Would you recommend rate plan migration as pro-active retention strategy?
  - The variables ovrmou_Mean, fe_vce_ovrrev_per has beta coefficients of 0.000248 and 0.00022 and is not a strong impact on churn. Hence rate plan migration as a pro-active strategy might not help.

Top Line Questions of Interest

- Recommendation on using this model to prioritize customers for pro-active retention strategy:
  - We could see from Gains table that top 20% of the probabilities lie between 0.321-0.805 and it contains 32.6% of the churn customers. As a pro-active strategy, we should target top 20 percentile of high churn probability customers using this model.
  - As an example, top 20% of the high churn probability customers were selected and written to csv file in the below path:

    C:\Jig19234\Capstone\target_top20percentile.csv

Customer Segments

- Customers were classified as High, Medium and Low based on average monthly revenue.
- Customers were classified as High, Medium and Low based on churn probability percent.
- We can target specific band of customers based on revenue and churn rate probability.
- As an example, High revenue and high probability rate customers were selected and written to csv file in the below path:

  C:\Jig19234\Capstone\target_seg.csv