TELECOM CHURN CASE STUDY

BY: AISHWARYA MALI

AMISH DESHPANDE

KONDAPALLI ANIL KUMAR

PROBLEM STATEMENT

- IN TELECOM INDUSTRY, CUSTOMERS ARE ABLE TO CHOOSE FROM MULTIPLE SERVICE PROVIDERS AND ACTIVELY SWITCH FROM ONE OPERATOR TO ANOTHER. IN THIS HIGHLY COMPETITIVE MARKET, THE TELECOMMUNICATIONS INDUSTRY EXPERIENCES AN AVERAGE OF 15-25% ANNUAL CHURN RATE.

- FOR MANY INCUMBENT OPERATORS, RETAINING HIGH PROFITABLE CUSTOMERS IS THE NUMBER ONE BUSINESS GOAL.

- TO REDUCE CUSTOMER CHURN, TELECOM COMPANIES NEED TO PREDICT WHICH CUSTOMERS ARE AT HIGH RISK OF CHURN.

- IN THIS PROJECT, WE WILL ANALYSE CUSTOMER-LEVEL DATA OF A LEADING TELECOM FIRM, BUILD PREDICTIVE MODELS TO IDENTIFY CUSTOMERS AT HIGH RISK OF CHURN AND IDENTIFY THE MAIN INDICATORS OF CHURN.

UNDERSTANDING AND DEFINING CHURN

- THERE ARE TWO MAIN MODELS OF PAYMENT POST-PAID AND PREPAID.

- IN THE POST-PAID MODEL, WHEN CUSTOMERS WANT TO SWITCH TO ANOTHER OPERATOR, THEY INFORM THE EXISTING OPERATOR TO TERMINATE THE SERVICES.

- IN THE PREPAID MODEL, CUSTOMERS WHO WANT TO SWITCH CAN SIMPLY STOP USING THE SERVICES, AND IT IS HARD TO KNOW WHETHER SOMEONE HAS ACTUALLY CHURNED OR IS SIMPLY NOT USING THE SERVICES TEMPORARILY.

- THUS, CHURN PREDICTION IS USUALLY MORE CRITICAL (AND NON-TRIVIAL) FOR PREPAID CUSTOMERS, AND THE TERM 'CHURN' SHOULD BE DEFINED CAREFULLY.

UNDERSTANDING BUSINESS OBJECTIVE AND THE DATA

- THE DATASET CONTAINS CUSTOMER-LEVEL INFORMATION FOR A SPAN OF FOUR CONSECUTIVE MONTHS. THE MONTHS ARE ENCODED AS 6, 7, 8 AND 9, RESPECTIVELY.

- THE BUSINESS OBJECTIVE IS TO PREDICT THE CHURN IN THE LAST MONTH USING THE DATA. TO DO THIS TASK, UNDERSTANDING THE TYPICAL CUSTOMER BEHAVIOUR DURING CHURN IS IMPORTANT.

- THERE ARE TWO TYPES OF CHURNS A)REVENUE-BASED CHURN AND B)USAGE-BASED CHURN.

UNDERSTANDING CUSTOMER BEHAVIOUR DURING CHURN

- CUSTOMERS USUALLY DO NOT DECIDE TO SWITCH TO ANOTHER COMPETITOR INSTANTLY, BUT RATHER OVER A PERIOD OF TIME. IN CHURN PREDICTION, WE ASSUME THAT THERE ARE THREE PHASES OF CUSTOMER LIFECYCLE:

    A)THE 'GOOD' PHASE

    B)THE 'ACTION' PHASE

    C)THE 'CHURN' PHASE

STEPS TO BUILD THE MODEL

- PRE-PROCESS DATA (CONVERT COLUMNS TO APPROPRIATE FORMATS, HANDLE MISSING VALUES, ETC.)

- CONDUCT APPROPRIATE EXPLORATORY ANALYSIS TO EXTRACT USEFUL INSIGHTS (WHETHER DIRECTLY USEFUL FOR BUSINESS OR FOR EVENTUAL MODELLING/FEATURE ENGINEERING).

- DERIVE NEW FEATURES.

- REDUCE THE NUMBER OF VARIABLES USING PCA.

- TRAIN A VARIETY OF MODELS, TUNE MODEL HYPERPARAMETERS, ETC. (HANDLE CLASS IMBALANCE USING APPROPRIATE TECHNIQUES).

- EVALUATE THE MODELS USING APPROPRIATE EVALUATION METRICS. NOTE THAT IT IS MORE IMPORTANT TO IDENTIFY CHURNERS THAN THE NON-CHURNERS ACCURATELY - CHOOSE AN APPROPRIATE EVALUATION METRIC WHICH REFLECTS THIS BUSINESS GOAL.

- CHOOSE A MODEL BASED ON SOME EVALUATION METRIC.

- FINALLY, RECOMMEND STRATEGIES TO MANAGE CUSTOMER CHURN BASED ON OUR OBSERVATIONS.

READING AND UNDERSTANDING THE DATA

-

|	mobile_number	circle_id	loc_og_t2o_mou	std_og_t2o_mou	loc_ic_t2o_mou	last_date_of_month_6	last_date_of_month_7	last_date_of_month_8	last_date_of
0	7000842753	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	
1	7001865778	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	
2	7001625959	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	
3	7001204172	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	
4	7000142493	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	

- FILTER HIGH-VALUE CUSTOMERS

- HANDLING MISSING VALUES IN COLUMNS

- HANDLING MISSING VALUES IN ROWS

- TAG CHURNERS

- OUTLIERS TREATMENT
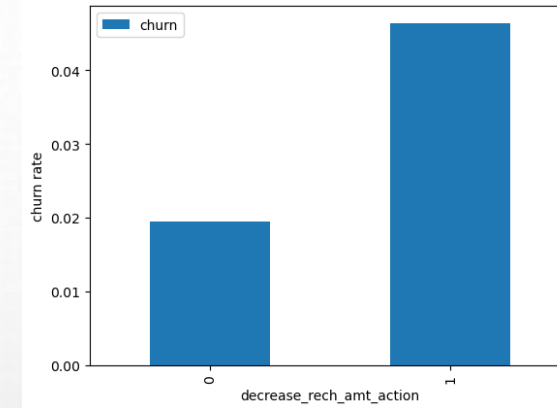
EXPLORATORY DATA ANALYSIS
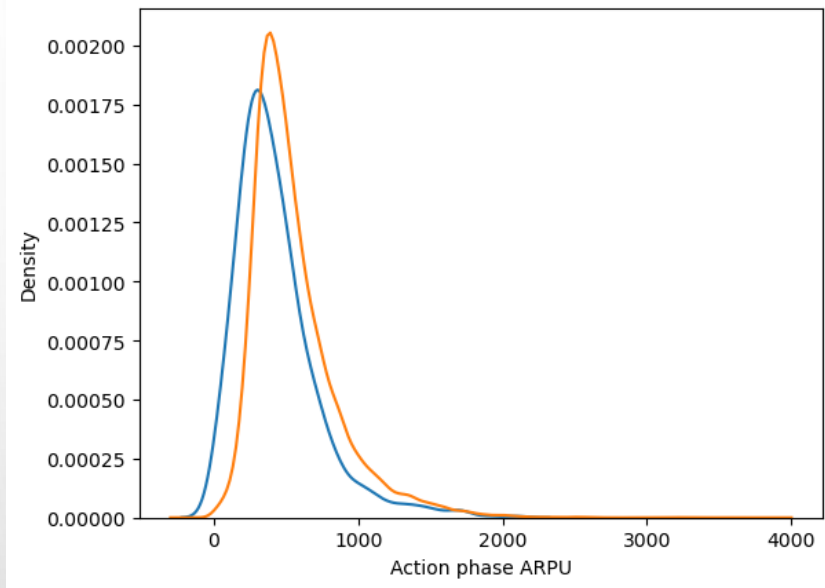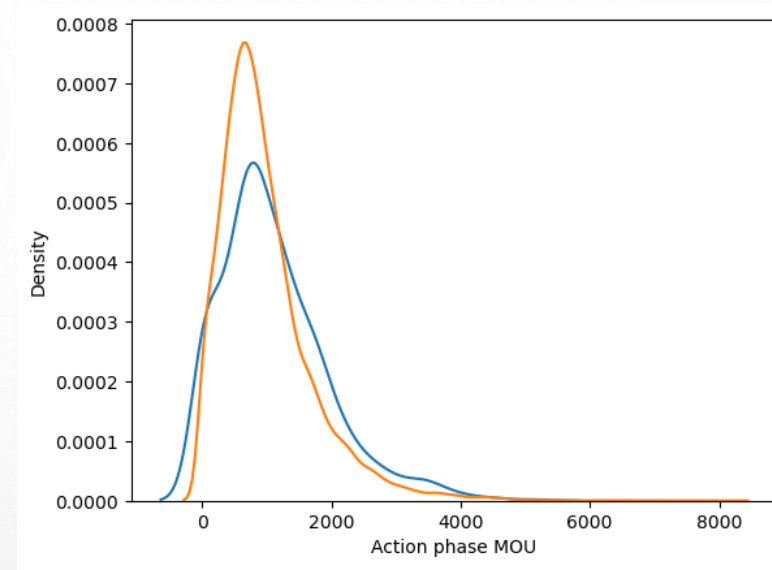
- **UNIVARIATE ANALYSIS**
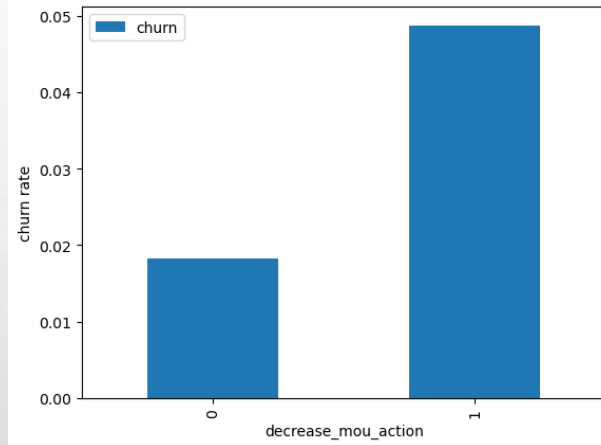
1
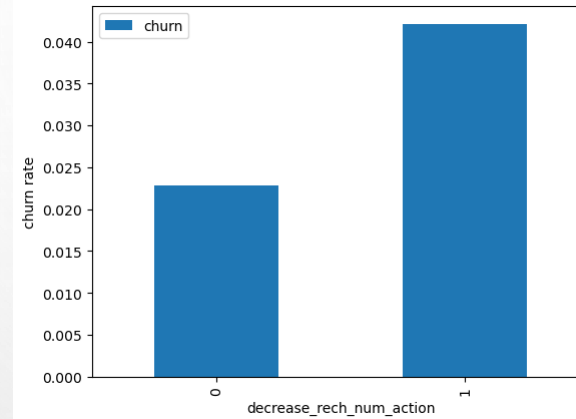
2

3

- 1) WE CAN SEE THAT THE CHURN RATE IS MORE FOR THE CUSTOMERS, WHOSE MINUTES OF USAGE(MOU) DECREASED IN THE ACTION PHASE THAN THE GOOD PHASE

- 2) AS EXPECTED, THE CHURN RATE IS MORE FOR THE CUSTOMERS, WHOSE NUMBER OF RECHARGE IN THE ACTION PHASE IS LESSER THAN THE NUMBER IN GOOD PHASE

- 3)HERE ALSO WE SEE THE SAME BEHAVIOUR. THE CHURN RATE IS MORE FOR THE CUSTOMERS, WHOSE AMOUNT OF RECHARGE IN THE ACTION PHASE IS LESSER THAN THE AMOUNT IN GOOD PHASE.

1   2
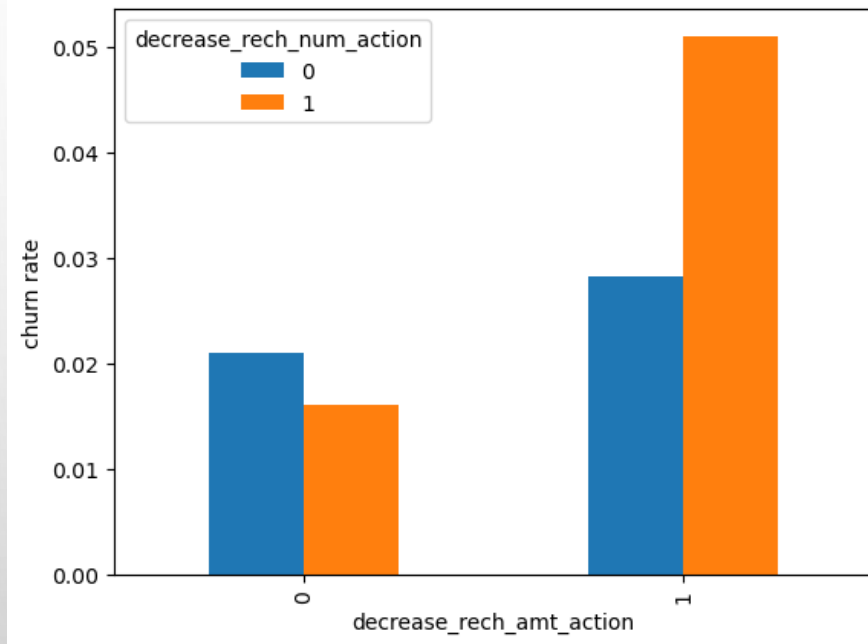
1) Average revenue per user (ARPU) for the churned customers is mostly densed on the 0 to 900. The higher ARPU customers are less likely to be churned.
ARPU for the not churned customers is mostly densed on the 0 to 1000.

2) Minutes of usage(MOU) of the churn customers is mostly populated on the 0 to 2500 range. Higher the MOU, lesser the churn probability.

BIVARIATE ANALYSIS
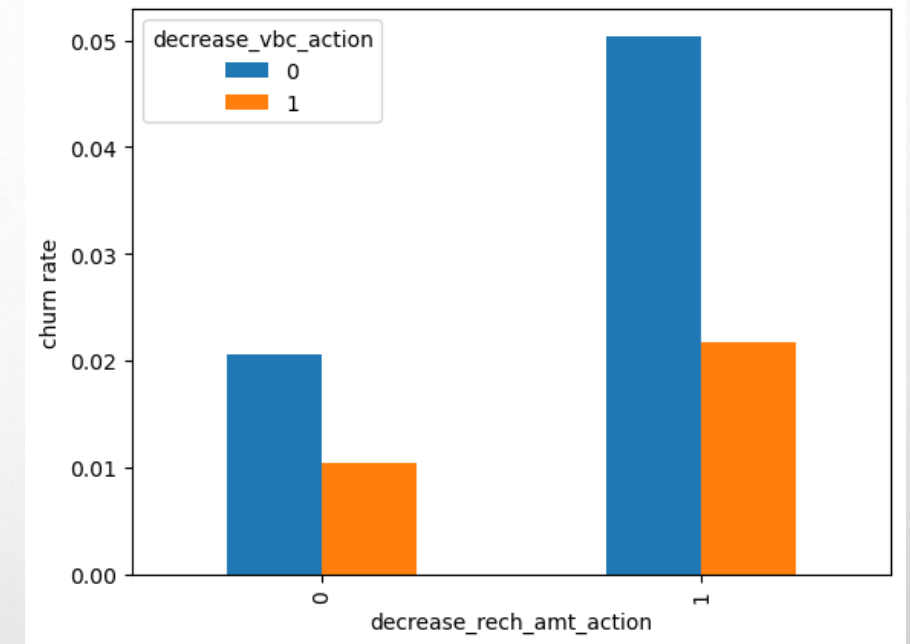
- 1) WE CAN SEE FROM THE ABOVE PLOT, THAT THE CHURN RATE IS MORE FOR THE CUSTOMERS, WHOSE RECHARGE AMOUNT AS WELL AS NUMBER OF RECHARGE HAVE DECREASED IN THE ACTION PHASE THAN THE GOOD PHASE.

- WE CAN SEE FROM THE ABOVE PLOT, THAT THE CHURN RATE IS MORE FOR THE CUSTOMERS, WHOSE RECHARGE AMOUNT AS WELL AS NUMBER OF RECHARGE HAVE DECREASED IN THE ACTION PHASE THAN THE GOOD PHASE.

PLOT OF C VERSUS TRAIN AND VALIDATION SCORES
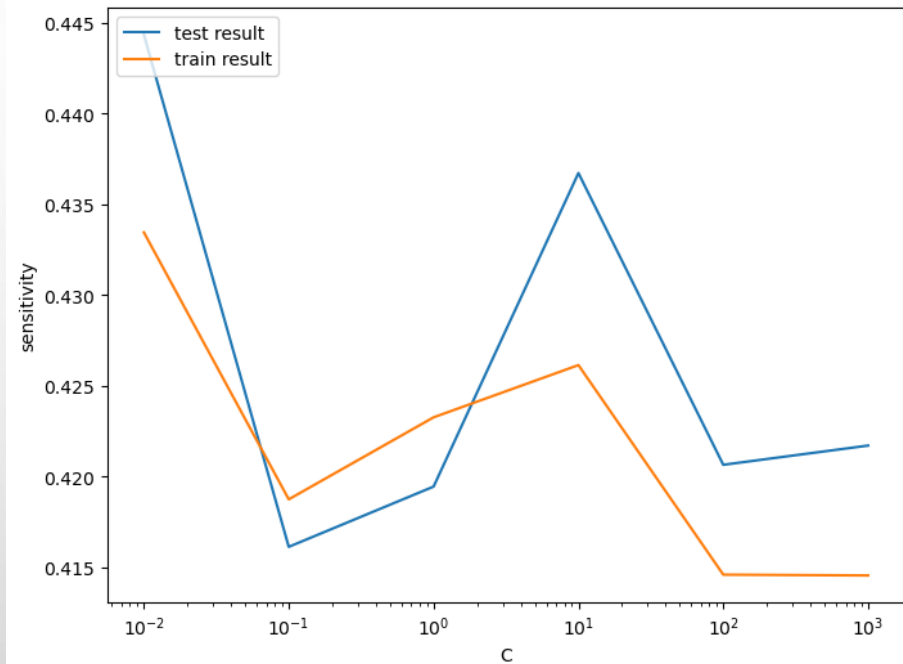
Accuracy:- 0.8751127955242736
Sensitivity:- 0.4167794316644114
Specificity:- 0.8909218203033838

Model summary

    Train set

Accuracy = 0.86

Sensitivity = 0.89

Specificity = 0.83

    Test set

Accuracy = 0.83

Sensitivity = 0.81

Specificity = 0.83

Overall, the model is performing well in the test set, what it had learnt from the train set.

MODEL BUILDING

- LOGISTIC REGRESSION

- FEATURE SELECTION USING RFE

- MODEL-1 WITH RFE SELECTED COLUMNS

RFE SELECTED COLUMNS

76637 2.157237e-02

44417 1.367050e-02

77582 4.069784e-04

48224 3.781698e-03

76321 7.378413e-14

ANALYSIS

Accuracy - Becomes stable around 0.6
Sensitivity - Decreases with the increased probablity.
Specificity - Increases with the increasing probablity.
At point 0.6 where the three parameters cut each other, we can see that there is a balance bethween sensitivity and specificity with a good accuracy.
Here we are intended to acheive better sensitivity than accuracy and specificity. Though as per the above curve, we should take 0.6 as the optimum probability cutoff, we are taking **0.5** for acheiving higher sensitivity, which is our main goal.

TESTING THE MODEL

MODEL SUMMARY

- TRAIN SET
  - ACCURACY = 0.84
  - SENSITIVITY = 0.81
  - SPECIFICITY = 0.83

- TEST SET
  - ACCURACY = 0.78
  - SENSITIVITY = 0.82
  - SPECIFICITY = 0.78

- OVERALL, THE MODEL IS PERFORMING WELL IN THE TEST SET, WHAT IT HAD LEARNT FROM THE TRAIN SET.

FINAL CONCLUSION WITH NO PCA

- WE CAN SEE THAT THE LOGISTIC MODEL WITH NO PCA HAS GOOD SENSITIVITY AND ACCURACY, WHICH ARE COMPARABLE TO THE MODELS WITH PCA. SO, WE CAN GO FOR THE MORE SIMPLISTIC MODEL SUCH AS LOGISTIC REGRESSION WITH PCA AS IT EXPLIAINS THE IMPORTANT PREDICTOR VARIABLES AS WELL AS THE SIGNIFICANCE OF EACH VARIABLE. THE MODEL ALSO HELS US TO IDENTIFY THE VARIABLES WHICH SHOULD BE ACT UPON FOR MAKING THE DECISION OF THE TO BE CHURNED CUSTOMERS. HENCE, THE MODEL IS MORE RELEVANT IN TERMS OF EXPLAINING TO THE BUSINESS

BUSINESS RECOMMENDATION

- TARGET THE CUSTOMERS, WHOSE MINUTES OF USAGE OF THE INCOMING LOCAL CALLS AND OUTGOING ISD CALLS ARE LESS IN THE ACTION PHASE (MOSTLY IN THE MONTH OF AUGUST).

- TARGET THE CUSTOMERS, WHOSE OUTGOING OTHERS CHARGE IN JULY AND INCOMING OTHERS ON AUGUST ARE LESS.

- ALSO, THE CUSTOMERS HAVING VALUE BASED COST IN THE ACTION PHASE INCREASED ARE MORE LIKELY TO CHURN THAN THE OTHER CUSTOMERS. HENCE, THESE CUSTOMERS MAY BE A GOOD TARGET TO PROVIDE OFFER.

- CUTOMERS, WHOSE MONTHLY 3G RECHARGE IN AUGUST IS MORE, ARE LIKELY TO BE CHURNED.

- CUSTOMERS HAVING DECREASING STD INCOMING MINUTES OF USAGE FOR OPERATORS T TO FIXED LINES OF T FOR THE MONTH OF AUGUST ARE MORE LIKELY TO CHURN.

- CUTOMERS DECREASING MONTHLY 2G USAGE FOR AUGUST ARE MOST PROBABLE TO CHURN.

- CUSTOMERS HAVING DECREASING INCOMING MINUTES OF USAGE FOR OPERATORS T TO FIXED LINES OF T FOR AUGUST ARE MORE LIKELY TO CHURN.

- ROAM_OG_MOU_8 VARIABLES HAVE POSITIVE COEFFICIENTS (0.7135). THAT MEANS FOR THE CUSTOMERS, WHOSE ROAMING OUTGOING MINUTES OF USAGE IS INCREASING ARE MORE LIKELY TO CHURN.