Customer Churn Prediction

Capstone Project

Vikash Kumar

DSBA – Aug'23

1

Business Problem Understanding

Problem, Objective, Scope and Constraints

BUSINESS PROBLEM & OBJECTIVE

- DTH provider facing heavy competition resulting in **customer churn**
- Loss of customers =Loss of company reputation=loss of bottom line= = Loss of Revenue
- Huge initial cost /customer & fixed content fee to broadcasters impacts profits
- Customer acquisition cost = 6 X customer retention cost
- Increasing customer retention by 5% increases profits from 25-95%
- Customer churn impacts topline and bottomline revenue

OBJECTIVE

The Churn rate is around 17% which is high. Hence, reducing the Churn to 10% will increase the revenue by 7%.

Predict customer churn so that segmented offers can be given as part of a retention campaign

SCOPE AND CONSTRAINTS

Scope

- Best performing model for churn prediction
- Key insights and recommendations from EDA and model

Constraints

- Focus of prior retention campaigns not known – cashbacks, coupons
- Campaign Budget not known

2

Data & Modelling

Modelling approach, model performance, best model

KEY INFORMATION ABOUT DATA

Shape

Dataset contains **11260 rows** and **19 columns**

- 5 float
- 2 integer
- 12 object

Nulls

- 4361 total nulls in all predictor fields
- **2.28%** of all predictor fields

Duplicates

0 duplicates in the dataset

Outliers

- 2658 outliers in numeric continuous columns
- Constitutes **1.4%** of all predictor fields

Data clean up

-**10 attributes** required clean-up
- Junk characters such as #, &, +, $, @ present
- Different representations of same category present. E.g., Male and M, Female and F

Target variable

-Churn = 1 (Churned customer)
Churn = 0 (Active customer)
-**16.8%** churned customers in dataset
- Class imbalance

Modelling Approach

Visual inspection
- Observe rows & columns, understand attributes, datatypes, nulls
- Perform data cleaning
- Statistical description

EDA
- Perform univariate and bivariate analysis
- Bad Data correction
- Correlations between variables

Data Preparation
- Missing value treatment
- Outlier treatment
- Variable transformation (Encoding)

Pre-processing
- Low VIF variables hence no major issue of multicollinearity in the model.
- Train-test split in 70:30 ratio
- Scaling

Modelling
- Try out different algorithms (30) & evaluate performance on train and test
- Benchmark base model performance
- Tune hyper parameters and change data
- Select best model based on evaluation metrics

Model evaluation criteria – Good fit, Interpretable, F1score, Precision, Recall for 1s

8 Algorithms selected
Classifier versions from SKLearn .

Base model with default hyperparameters for benchmarking

Performance improvement
1. Data changed for different trials
2. Hyperparameters tuned (Grid Search)
3. Regularization
4. Ensembles

Best model within each algorithm selected based on evaluation criteria

Model performance comparison across best models of all algorithms

Decide best model

COMPARISON OF MODEL PERFORMANCE

TOP performing Models Summary							
Model	**Hyperparameter**	**DataSet**	**Accuracy Score**	**Precision Score**	**Recall**	**AUC score**	**F1 Score**
RandomForestClassifierBag	Base Model of Bagging	Test	**0.99**	**0.97**	**0.99**	**0.99**	**0.99**
		Train	**1**	**1**	**1**	**1**	**1**
RandomForestClassifier	Base Model	Test	**0.99**	**0.97**	**0.99**	**0.99**	**0.99**
		Train	**1**	**1**	**1**	**1**	**1**
DecisionTreeClassifierGS	GridsSearch CV , Best model for F1 score	Test	**0.99**	**0.97**	**0.97**	**0.98**	**0.99**
		Train	**1**	**1**	**1**	**1**	**1**
DecisionTreeClassifier	Base Model	Test	**0.99**	**0.97**	**0.97**	**0.98**	**0.99**
		Train	**1**	**1**	**1**	**1**	**1**
RandomForestClassifierRS	Resampling Model, Best model for F1 score	Test	**0.96**	**0.93**	**1**	**0.96**	**0.96**
		Train	**1**	**1**	**1**	**1**	**1**
RandomForestClassifierRS	Resampling Model, Best model for F1 score	Test	**0.96**	**0.92**	**1**	**0.96**	**0.96**
		Train	**1**	**1**	**1**	**1**	**1**
DecisionTreeClassifierRS	Resampling Model, Best model for F1 score	Test	**0.96**	**0.92**	**0.99**	**0.96**	**0.96**
		Train	**1**	**1**	**1**	**1**	**1**
GaussianNBRS	Resampling Model, Best model for F1 score	Test	**0.96**	**0.93**	**0.99**	**0.96**	**0.96**
		Train	**1**	**1**	**1**	**1**	**1**

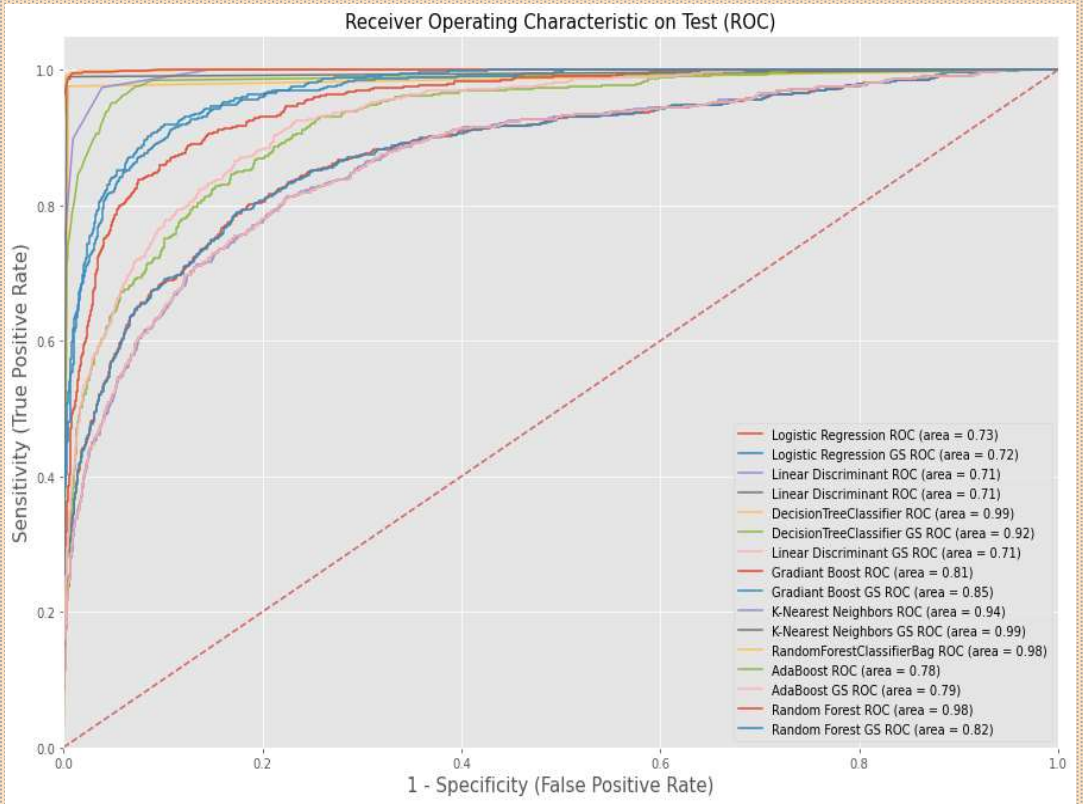
Detailed model performance of Gradient boost is given in Appendix

Best model

Random Forest *(Bagging)* is the best model! Why?

ROC curve and AUC

ROC : Performance of classification model at all classification thresholds
AUC: Signifies ability of model to differentiate between 0s and 1s



Where
True positive rate/ Sensitivity/Recall = TP/ TP+FN = TP/ Actual total positives
False positive rate/1-Specificity = FP/FP+TN = FP/ Actual total negatives

Why Gradient Boost?

Good fit

Comparable train and test performances

Precision, F1 score & Accuracy

Precision, Recall, F1-Score, Accuracy and AUC are highest

Interpretable

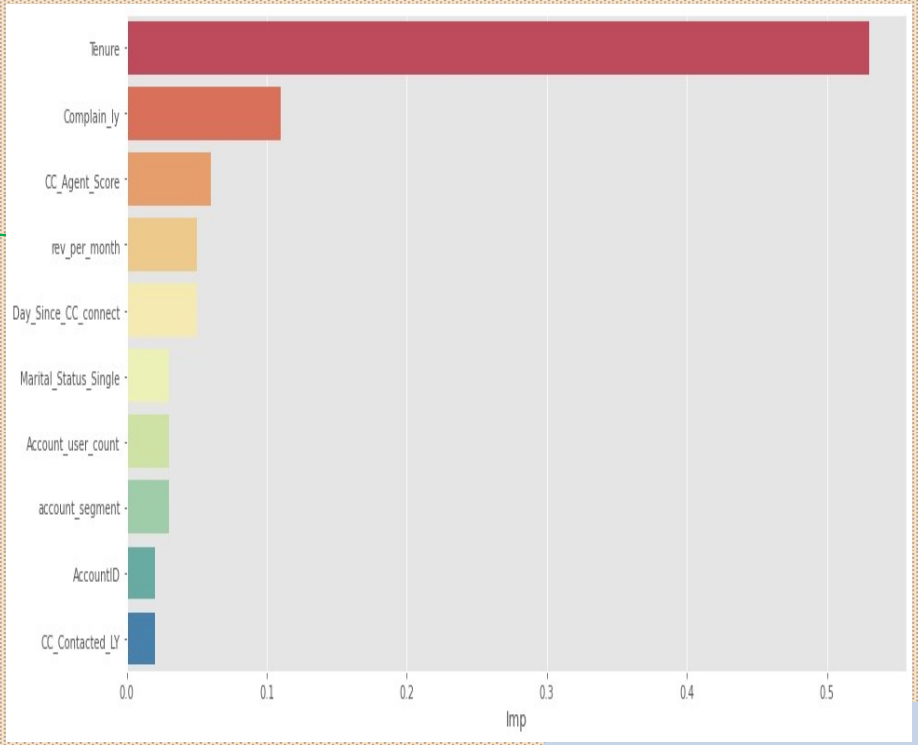
Random forest Feature importances provided by Sklearn

Top 10 predictors for the best model

- **Tenure** has the highest influence on the model – contributes 53% of all attributes feature importance. Low tenure (<2) has the highest churn
- The next 4 features together contributes to 25% of importance
  - **Days since customer care contact** - 5%
  - **Number of times customer care** was contacted last year – 11%
  - Account User Count– 3%
  - **Customer care agent score** – 6%

These are all related to **customer care and service**. Points to scope for improvement

- **Revenue per month** attribute contributes to 5% of all feature importance. The churn in high revenue customers is more than in low revenue customers.
- Marietal Status and other features all have individual contributions of less than 3% each



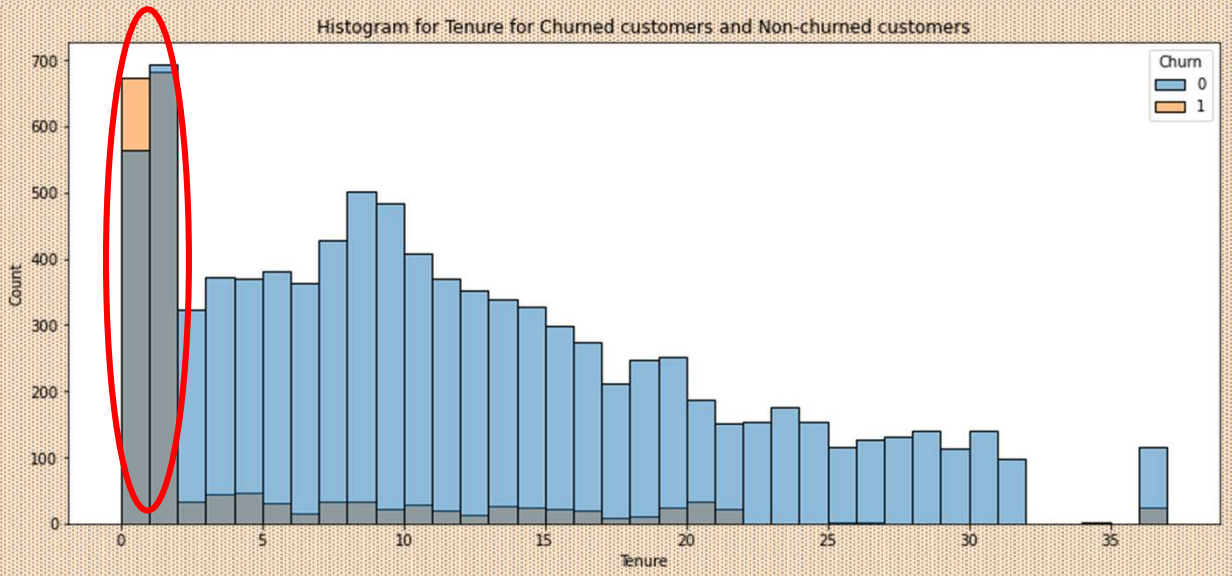
3

Key Business Insights & Recommendations

From Model interpretation and EDA

Click here for
- Detailed Recommendations
- Other Insights

High churn rate in low tenure customers

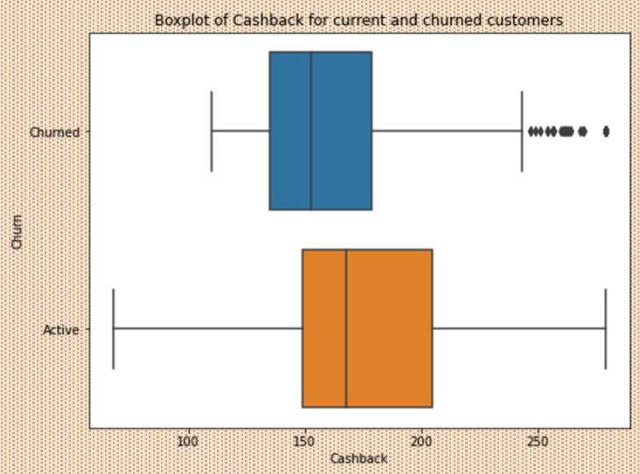


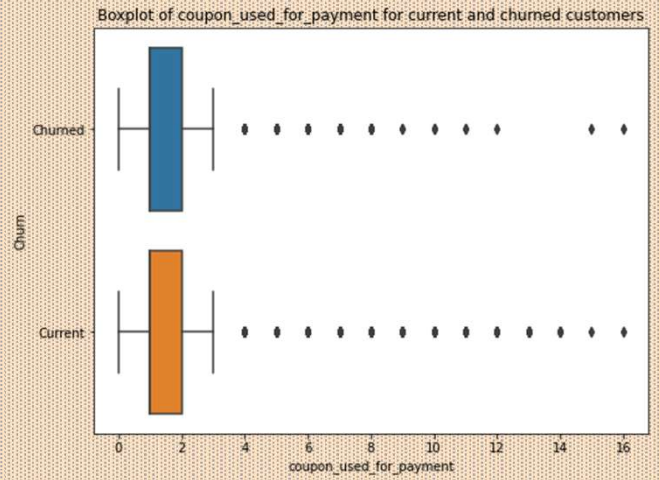
Insight: ***Churn is highest between the tenure period 0 to 2***

Recommendation: *There needs to be more focus initially as soon as customers are onboarded, to ensure they do not leave*

Activation/Onboarding team could **extend support** beyond the initial setup until customers settle down with the service
Customer care to initiate feedback survey to ensure customers have settled into the service

Cashback & Coupons





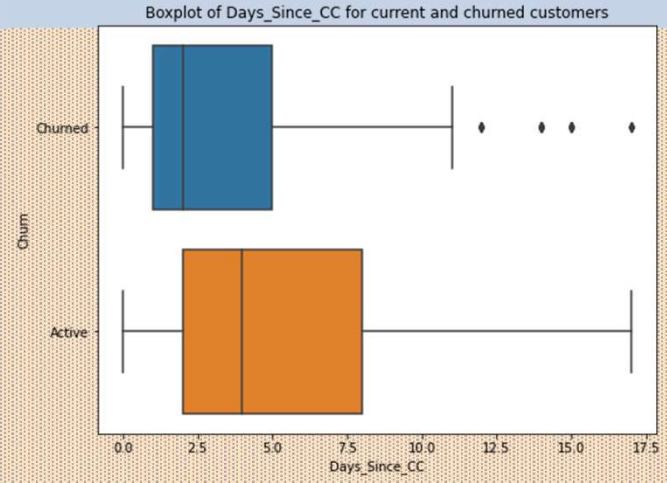
- The churned customers as shown in first boxplot have lesser cashback
- The churned and active customers have almost used the same number of coupons for payment

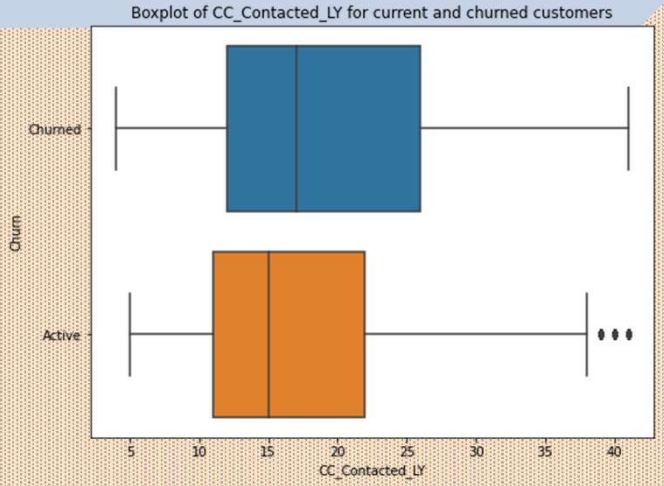
Insight : The current retention programs do not seem to be focusing on the customers with higher risk of churn

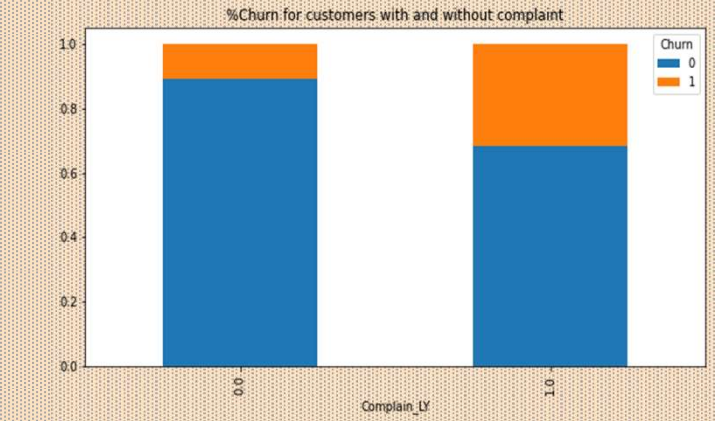
Recommendation: *Review whether existing cashback and coupon programs relevancy in current scenario , given the current churn model*

If they are not relevant, Need to design new retention programs to address current high risk customer group

Churn and Customer care service







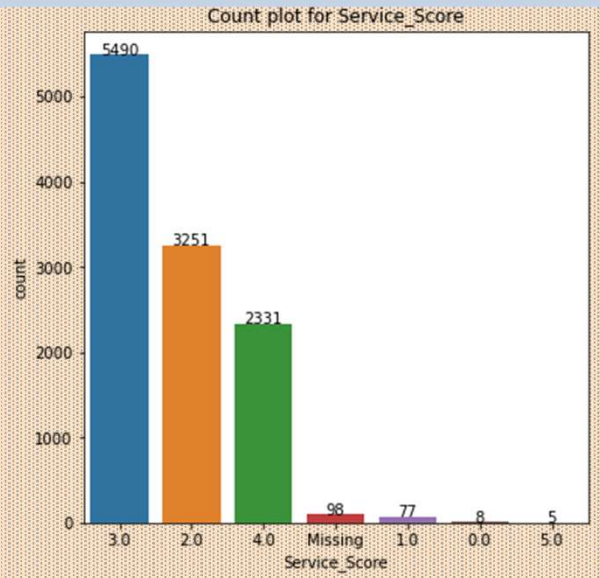
- Churned customers seem to have contacted customer care more recently before churning
- The number of times churned customers contacted customer care in the year is higher than number of times active customers contacted customer care
- 31% of customers who registered complaint churned Vs 11% of customers who have not registered complaint in last year

Insight: These indicate behavioral changes in customer before churn happens

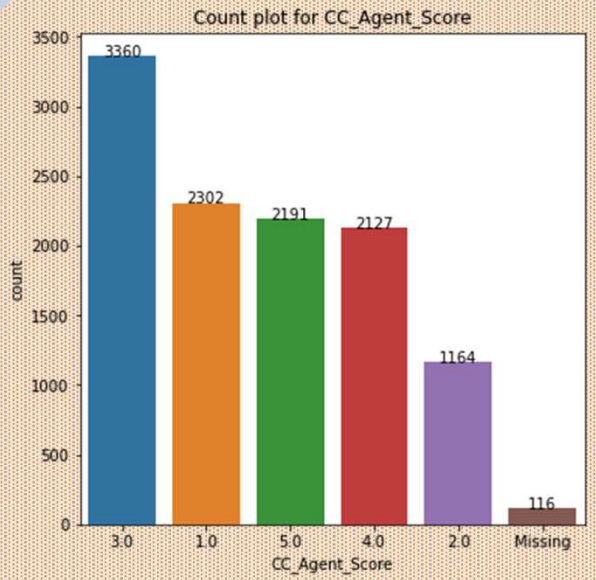
Recommendation: *Analyze Nature of Complaints & Customer care contact reasons*

- Need Root cause analysis to identify the key reason and fix top reasons
- Establish Service level payment agreements on basis of customer score (if not)

Customer care & Service – Customer perspective



Insight: 78% of customers have rated service as 3 or less than 3



Insight: 61% of customers have rated customer care agents a score of 3 or less than 3

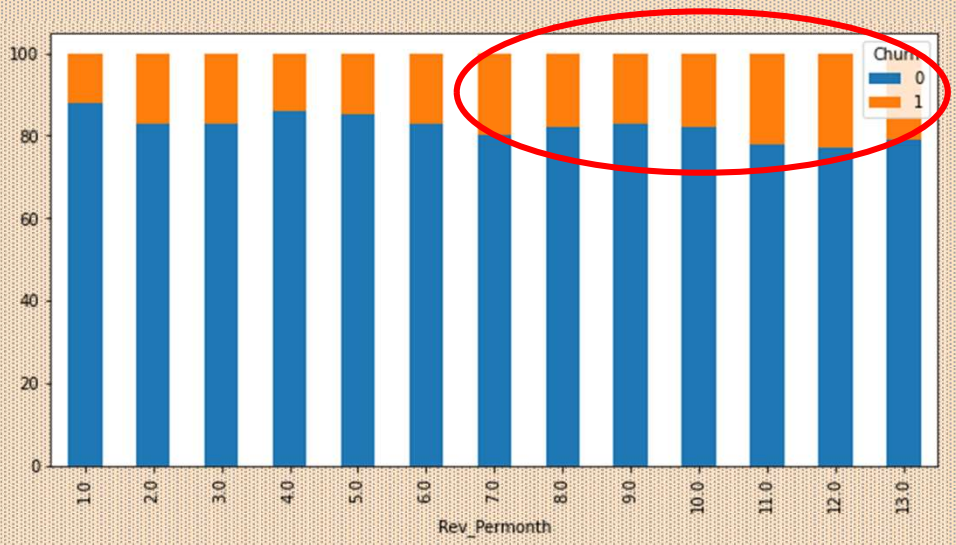
Recommendation: *Analyze customer feedback category wise nature of complaints*

Perform **Sentiment analysis** of the feedback (if any) that went along with scores

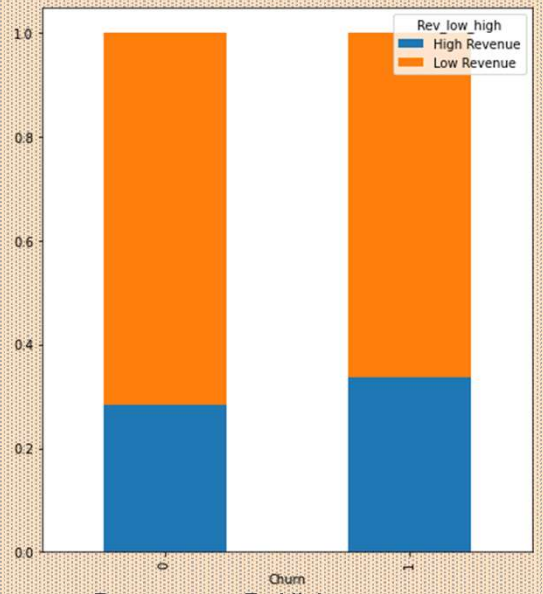
Identify top reasons that have resulted in low scores; if subjective feedback not captured, capture that as well

Revenue per month

Stacked bar chart showing
percent break up of active and churned customers by revenue



Stacked bar chart showing
percent break up of high and low revenue by customer status

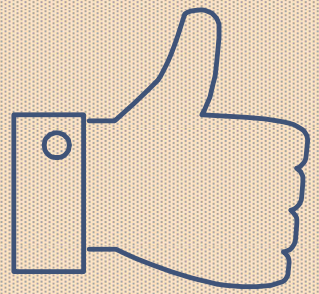


Revenue > = 7 : High revenue
Revenue < 7 : Low revenue

The % churn of customers in higher revenue group, for revenue >=7 per month is higher than low revenue group customers

Insight: More proportion of high revenue customers are leaving compared to less revenue customers

Recommendation: *Revisit the* segmented offers for high revenue customers



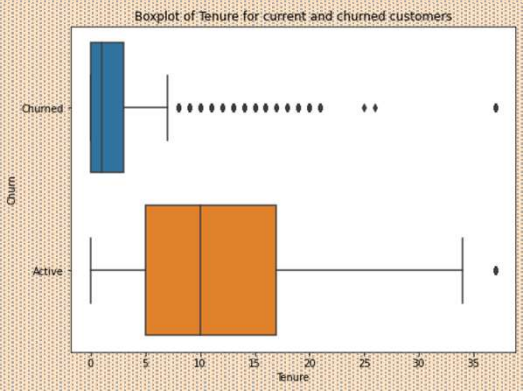
THANKS!

Appendix

Data Insights

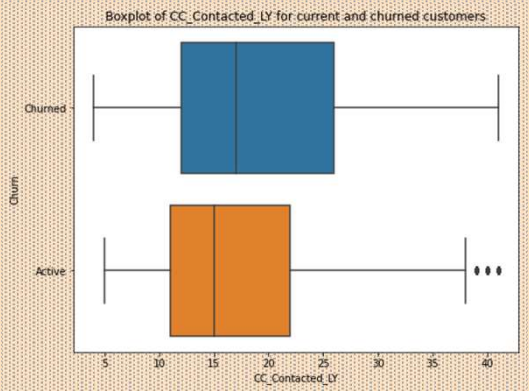
EDA plots

EDA – CONTINUOUS ATTRIBUTES



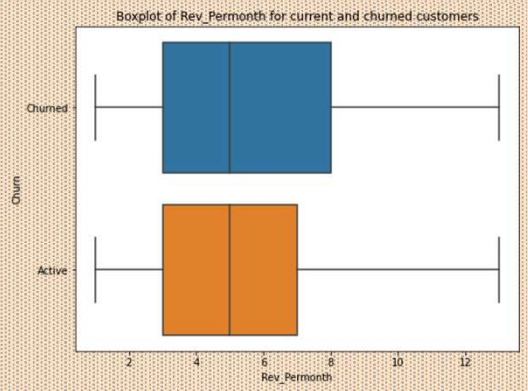
Tenure

The tenure of customers who have churned is much lesser than tenure of customers who are active



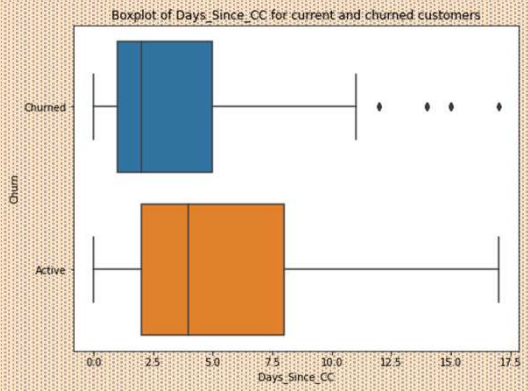
Times Customer Care contacted last year

Churned customers have contacted customer care more times than Active customers



Revenue per month

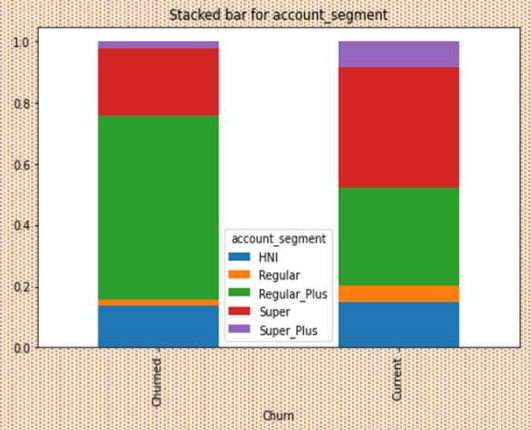
75th percentile of monthly revenues is higher in churned customers compared to active customers



Days since customer care contacted

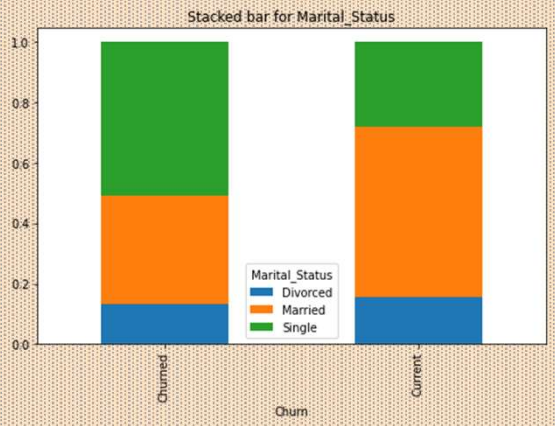
Customer churn seems to happen within few days of customer care contact (75 percentile – 5 days)

EDA – CATEGORICAL ATTRIBUTES



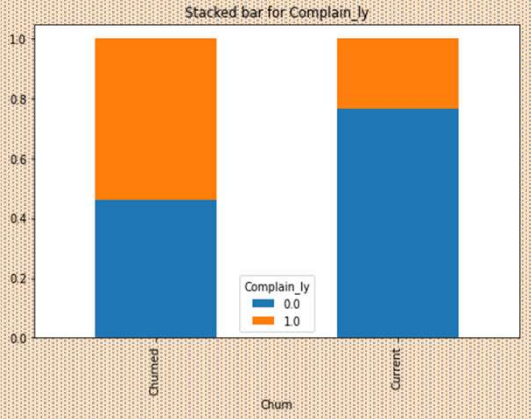
Account Segment

Compared to active customers, more Regular plus customers have churned



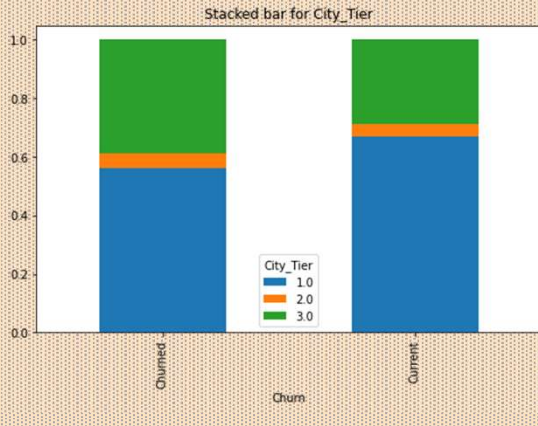
Marital Status

More single people show a propensity to churn compared to other status



Complaint last year?

More churned customers have registered complaints than active customers



City Tier

The proportion of Tier-3 customers in Churn is more than Active customers

Back

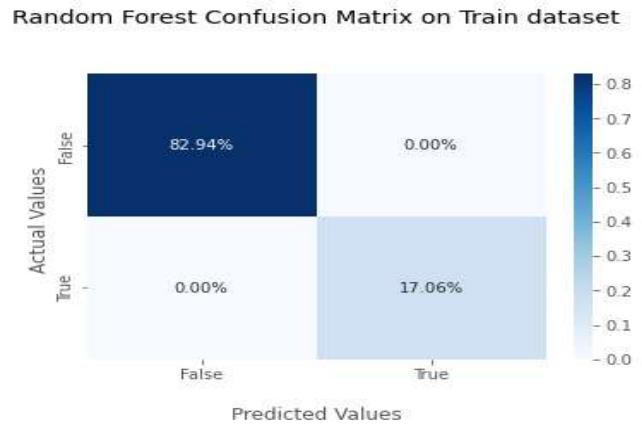
Model performance – Random Forest *(Bagging)*

Train data

Classification report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	7471
1	1.00	1.00	1.00	1537
accuracy			1.00	9008
macro avg	1.00	1.00	1.00	9008
weighted avg	1.00	1.00	1.00	9008

Confusion matrix

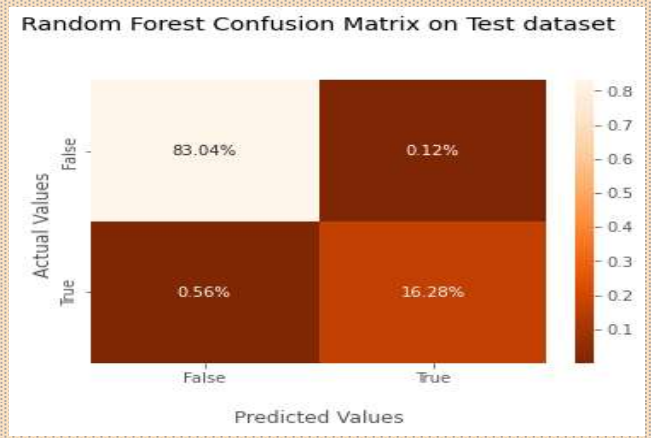


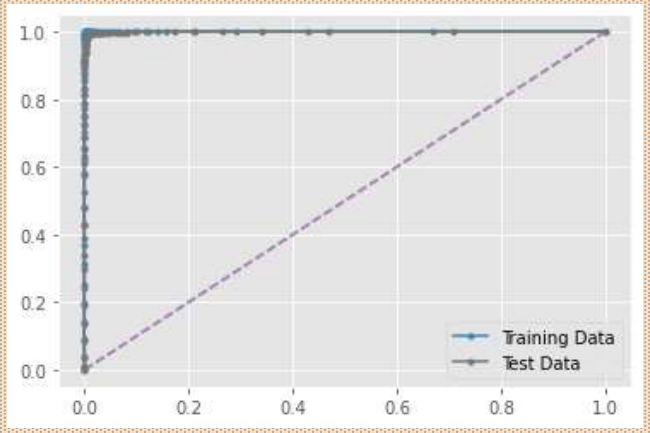
Test data

Classification report

	precision	recall	f1-score	support
0	0.99	1.00	1.00	2809
1	0.99	0.97	0.98	569
accuracy			0.99	3378
macro avg	0.99	0.98	0.99	3378
weighted avg	0.99	0.99	0.99	3378

Confusion matrix





5 - fold CV F1-score : **99**
Test data AUC score: 99
Test data F1-score : **99**

Test data performance

Precision = TP/TP+FP = 0.97
(True positives/ Predicted positives)

Recall = TP/TP+FN = 0.99
(True positives/ Total actual positives)

Back

DETAILED RECOMMENDATIONS

From EDA and Model

ONBOARDING/ACTIVATION PROCESS

Both EDA and Model have shown Tenure to have the highest impact on Churn

Possible Reasons: Bad first experience or Trial periods/prepaid accounts that expire automatically if no top-up is done within a predefined period.

- Important to determine between the above two reasons. Based on high customer care calls, complaints registered and low cashback and coupons for low tenure customers, it points to the first reason
- **Activation**/Onboarding team could **extend support** beyond the initial setup until customers settle down with the service
- Activation team **proactively engages customers** for the first month or two
- Customer care take a **feedback survey** about the process so that any hiccups can be understood and sorted out
- To increase response rates for feedback, gift cards/coupons can be given

COMPLAINT ANALYTICS AND MANAGEMENT

Complaint registered customers churned 182% times more than customers without complaint

Possible Reasons: Complaint not resolved on time or to customer's satisfaction.

- **Analyze complaints** from customers and identify top reasons
- Focus on **root cause** contributing to top reasons customers complain and address the root causes
- Establish **Service level agreements** to track and close complaints
- Ensure **proactive engagement through status updates** with customers until complaint is open
- Take **customer feedback** after complaint has been closed

REVIEW CUSTOMER CARE AND SERVICE PROCESS

78% of customers have rated service as 3 or less than 3 (out of a scale of 5)
61% of customers have rated customer care agents a score of 3 or less than 3 (out of 5)

Possible Reasons:

- Service score could be reflective of issues ranging from technical or plan structure or support from activation and customer care teams
- Customer care agent score may be reflective of agent behavior (e.g., knowledge, disposition) or process (e.g., call wait time)

- It is recommended to do a **root cause analysis** (pareto) of the feedback (if any) that went along with scores.
- Focus on root causes for top reasons for low feedback scores. There may be pointers to reasons for churn as well
- If verbal feedback not captured currently, it is suggested to change the process to capture that

REVIEW CURRENT RETENTION SPENDS

The median cashback is 10% more for active customers than churned customers
Same number of coupons have been used by churned and active customers

Possible Reasons: Existing retention programs may not specifically be targeting customers at risk for churn (or)
Existing programs may have addressed earlier high risk churn segment thereby reducing churn in that segment and hence current risk profile has changed

- **Review whether existing cashback and coupon programs are still relevant** given the current churn model
- If they are not relevant, design new retention programs to address current high risk customer group

Back

References

Special thanks to people who offered these resources for free:

- Presentation template by SlidesCarnival
- Photographs by Startup Stock Photos

REFERENCES

- Essentials of customer churn and retention | Smartlook Blog
- DTH industry: A glimpse of profits at last! | Business Standard News (business-standard.com)
- Customer Retention Marketing vs. Customer Acquisition Marketing | OutboundEngine