Capstone Project

Final Report

Name: Sahithi Pothuri

PGDSBA -Jan B batch 2022

Table of Contents PGNO

List of Figures pgno

List of Tables pgno

1) Introduction of the business problem

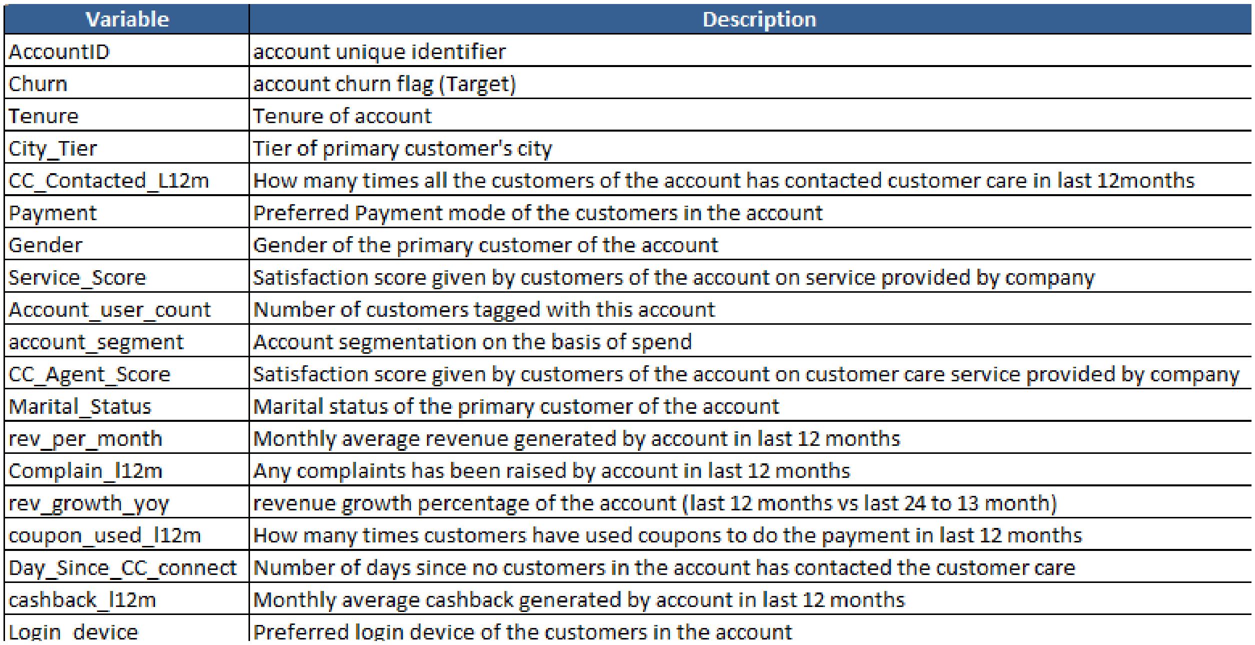
a) Defining problem statement

Given Business problem is that of an DTH provider who is facing a lot of competition in the current market and it has become a challenge to retain the existing customers in the current situation.

The given information is collected within time range of 12 months.

The problem statement aims at predicting churned customers.

Below is the data dictionary of the given data set



Variable	Description
AccountID	account unique identifier
Churn	account churn flag (Target)
Tenure	Tenure of account
City_Tier	Tier of primary customer's city
CC_Contacted_L12m	How many times all the customers of the account has contacted customer care in last 12months
Payment	Preferred Payment mode of the customers in the account
Gender	Gender of the primary customer of the account
Service_Score	Satisfaction score given by customers of the account on service provided by company
Account_user_count	Number of customers tagged with this account
account_segment	Account segmentation on the basis of spend
CC_Agent_Score	Satisfaction score given by customers of the account on customer care service provided by company
Marital_Status	Marital status of the primary customer of the account
rev_per_month	Monthly average revenue generated by account in last 12 months
Complain_l12m	Any complaints has been raised by account in last 12 months
rev_growth_yoy	revenue growth percentage of the account (last 12 months vs last 24 to 13 month)
coupon_used_l12m	How many times customers have used coupons to do the payment in last 12 months
Day_Since_CC_connect	Number of days since no customers in the account has contacted the customer care
cashback_l12m	Monthly average cashback generated by account in last 12 months
Login_device	Preferred login device of the customers in the account

1A

b) Business Objectives

The company needs unique, clear campaign suggestions and recommendations to provide segmented offers to the potential churners.

c) Business Constraints

The company's revenue assurance team expects recommendations to be profitable and will not be approved if they incur loss to the company.

The data given only deals with features related to customer experience, account transaction history and has provided minimum details on financial status and monetary transactions of customer.

b) Need of the study/project

In the given problem, the churned customers are to be identified.

- Acquiring a new customer can cost five times more than retaining an existing customer.
- Increasing customer retention by 5% can increase profits from 25-95%.
- The success rate of selling to a customer you already have is 60-70%, while the success rate of selling to a new customer is 5-20%.
- U.S. companies lose $136.8 billion per year due to avoidable consumer switching.
- American Express found 33% of customers will consider switching companies after just one instance of poor customer service.

Considering the above analysis, it is very profitable to retain existing customer rather than creating new costumer. The business constrains we have discussed above also stresses the importance of retaining customer as it is profitable. The given project helps the company to take proactive decisions to avoid customer churn and there by helps in customer segmentation, application of various market strategies on segmented customers.

2) Data Report

a) Understanding how data was collected in terms of time, frequency and methodology.

Time span

The given data is collected in a continuous time span of 12 months.

Revenue growth is also measured for 12 months Vs last 12 months.

Frequency of data

The data is collected randomly and thoroughly from various samples to ensure the final distribution of data follows normal distribution.

This step is crucial as sample data represents the population and gives accurate estimates of population.

Methodology used

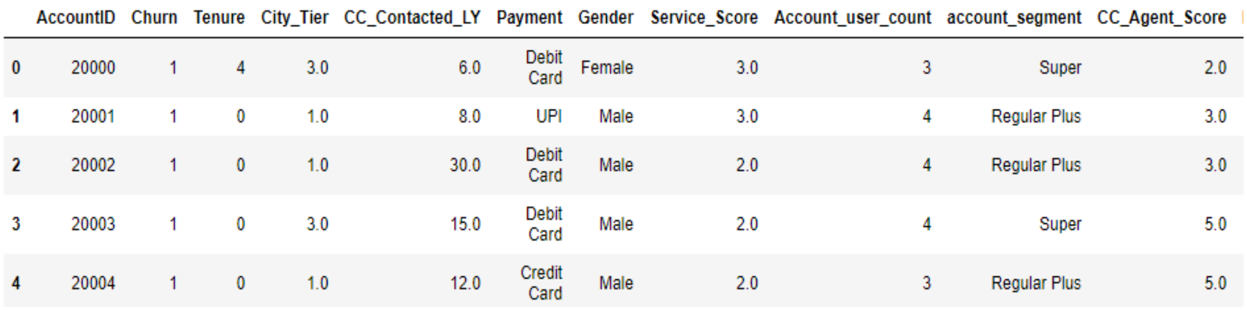
When the data in different entries and features are observed, it can be understood that data is collected from multiple sources/systems/databases, as there is mismatch in the categories within few features.

The mismatch is only observed in categories names but not in values, it infers that the data is collected through primary source that means collected by the researcher himself.

There is also chance of secondary source of data collected through surveys, feedback, forms, interviews, which the researcher merged well within primary source.

b) Visual inspection of data (rows, columns, descriptive details)

The dataset begins with account ID representing uniqueness of each entry, every entry is data collected from a customer.

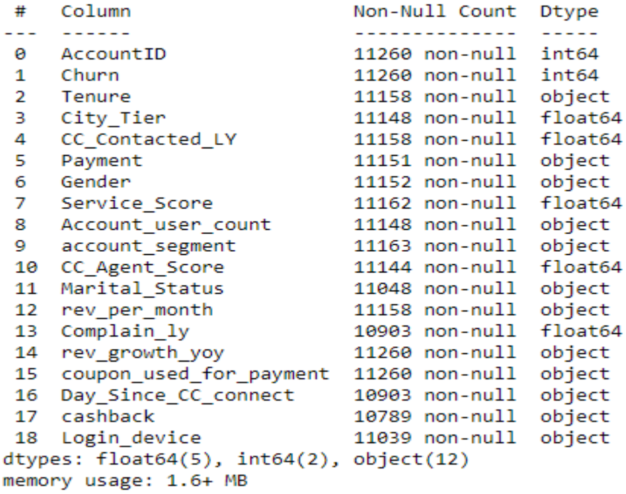


	AccountID	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score
0	20000	1	4	3.0	6.0	Debit Card	Female	3.0	3	Super	2.0
1	20001	1	0	1.0	8.0	UPI	Male	3.0	4	Regular Plus	3.0
2	20002	1	0	1.0	30.0	Debit Card	Male	2.0	4	Regular Plus	3.0
3	20003	1	0	3.0	15.0	Debit Card	Male	2.0	4	Super	5.0
4	20004	1	0	1.0	12.0	Credit Card	Male	2.0	3	Regular Plus	5.0

1B

There are 11260 entries and 19 features in the data.

Below are data types present in dataset



```
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   AccountID                11260 non-null  int64
 1   Churn                    11260 non-null  int64
 2   Tenure                   11158 non-null  object
 3   City_Tier                11148 non-null  float64
 4   CC_Contacted_LY          11158 non-null  float64
 5   Payment                  11151 non-null  object
 6   Gender                   11152 non-null  object
 7   Service_Score            11162 non-null  float64
 8   Account_user_count       11148 non-null  object
 9   account_segment          11163 non-null  object
 10  CC_Agent_Score           11144 non-null  float64
 11  Marital_Status           11048 non-null  object
 12  rev_per_month            11158 non-null  object
 13  Complain_ly              10903 non-null  float64
 14  rev_growth_yoy           11260 non-null  object
 15  coupon_used_for_payment  11260 non-null  object
 16  Day_Since_CC_connect     10903 non-null  object
 17  cashback                 10789 non-null  object
 18  Login_device             11039 non-null  object
dtypes: float64(5), int64(2), object(12)
memory usage: 1.6+ MB
```

1C

There are 5 float data types, there are two integer data types, both are categorical (Account ID, customer churn) and there are 12 object data types. There is mismatch in data types, object data types need to be encoded and there may be change in count of data types after Preprocessing and encoding.

Null values

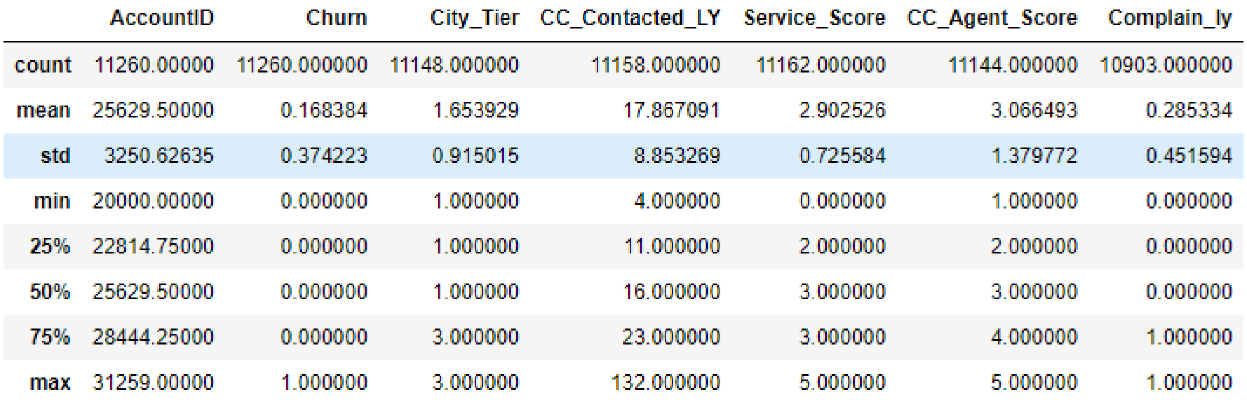
There are null values in every feature; these null values exist due to lack of information in a particular feature for an entry.

If null values in a feature are greater than 60%, they need to be dropped else these null values need to be imputed with proper measure of central tendency like mean, median and mode.

Duplicates values

There are no duplicate values in the data set, it can be inferred that the researcher didn't attempt to use existing entries in place of missing entries.

Statistical summary of numeric data type

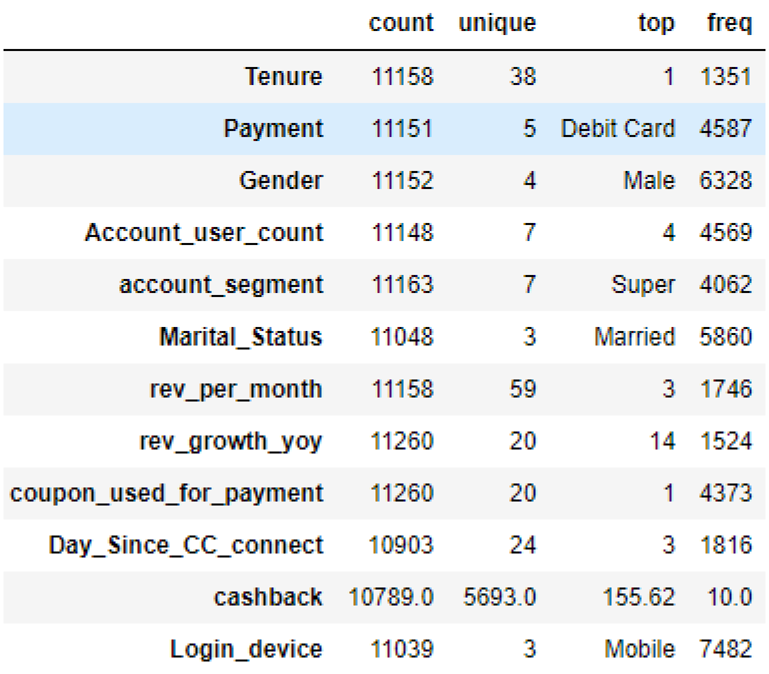


	AccountID	Churn	City_Tier	CC_Contacted_LY	Service_Score	CC_Agent_Score	Complain_ly
count	11260.00000	11260.000000	11148.000000	11158.000000	11162.000000	11144.000000	10903.000000
mean	25629.50000	0.168384	1.653929	17.867091	2.902526	3.066493	0.285334
std	3250.62635	0.374223	0.915015	8.853269	0.725584	1.379772	0.451594
min	20000.00000	0.000000	1.000000	4.000000	0.000000	1.000000	0.000000
25%	22814.75000	0.000000	1.000000	11.000000	2.000000	2.000000	0.000000
50%	25629.50000	0.000000	1.000000	16.000000	3.000000	3.000000	0.000000
75%	28444.25000	0.000000	3.000000	23.000000	3.000000	4.000000	1.000000
max	31259.00000	1.000000	3.000000	132.000000	5.000000	5.000000	1.000000

1D

The above is the table which describes the five point summary of the data with min, max, 1st quartile , 3rd quartiles , standard deviation and mean in a feature. It can be observed there is no huge difference between max and 75% of data inferring less number of outliers.

Statistical summary of object data type



	count	unique	top	freq
Tenure	11158	38	1	1351
Payment	11151	5	Debit Card	4587
Gender	11152	4	Male	6328
Account_user_count	11148	7	4	4569
account_segment	11163	7	Super	4062
Marital_Status	11048	3	Married	5860
rev_per_month	11158	59	3	1746
rev_growth_yoy	11260	20	14	1524
coupon_used_for_payment	11260	20	1	4373
Day_Since_CC_connect	10903	24	3	1816
cashback	10789.0	5693.0	155.62	10.0
Login_device	11039	3	Mobile	7482

1E

The above table provides details of top category or data point with frequency and count. Debit card payment is more preferred, while most of the customers are male, married, most used device is Mobile with top account segment super.

c) Understanding of attributes (variable info, renaming if required)

Attributes are variables which creates memory for data, the data can be either qualitative or quantitative, where qualitative is countable and quantitative is measurable.

The given problem statement is a combination of both qualitative and quantitative data like Gender, marital status, revenue, revenue growth , satisfaction score, churn etc.

Few variables are misinterpreted as another data type irrespective of their own type, which needs to be thoroughly checked and preprocessed.

Renaming of variables is not needed for the given problem statement and all variable names are meaningful and easy to understand.

3. Data Cleaning and Pre-processing

There are many missing values within each feature, the data set need to be thoroughly preprocessed in order to get best insights.

Approach used for identifying missing Values:

Below approaches used are aimed to create clean data set with no null values, special categories, repetitive category names.

Frequency of every unique data point and category is observed for both Categorical and numerical data types using value counts.

Similar Category with different name like 'Male' and 'M' are replaced with only one Category name.

There are certain features like cash back, where special character are identified by thorough analysis of entire columns.

a)Pre-processing:

1. There are special characters in few features which have been replaced with new category.
2. Null values have been replaced with suitable measures of central tendency like mean, median, mode.
3. Categorical data types are replaced with mode
4. Numerical data types are replaced with median.

Approach used for identifying Outliers:

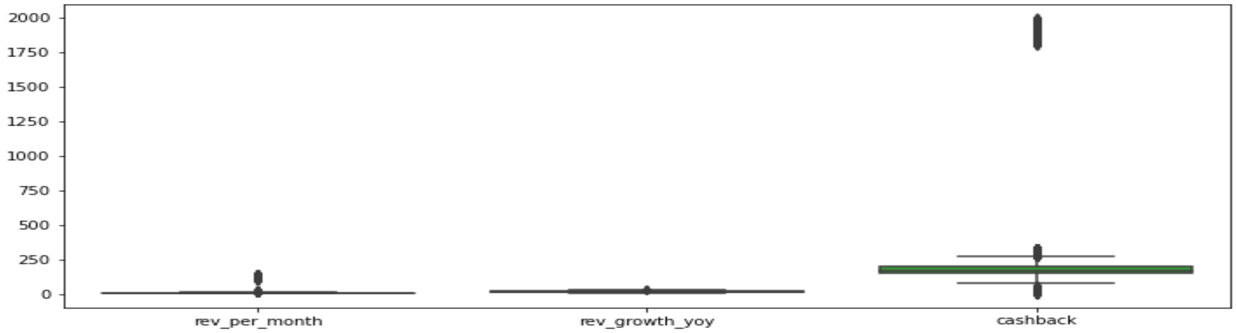
Box plots are used to identify outliers for numerical data types, outliers observed beyond upper whisker and below lower whisker are imputed with Q1-1.5*IQR, Q3+1.5*IQR and are not dropped.

b)Outlier treatment (if required)

Outliers are data points which are dispersed away from other group of data points, these outliers makes the distribution skewed.

Outlier removal is based on the distribution within each feature and need not be same for all features. Outlier treatment can only be done on continuous data types, there are three features which are continuous - rev_per_month, rev_growth_yoy, cashback.

Below is the Box plot visualizing outliers in each feature mentioned above.

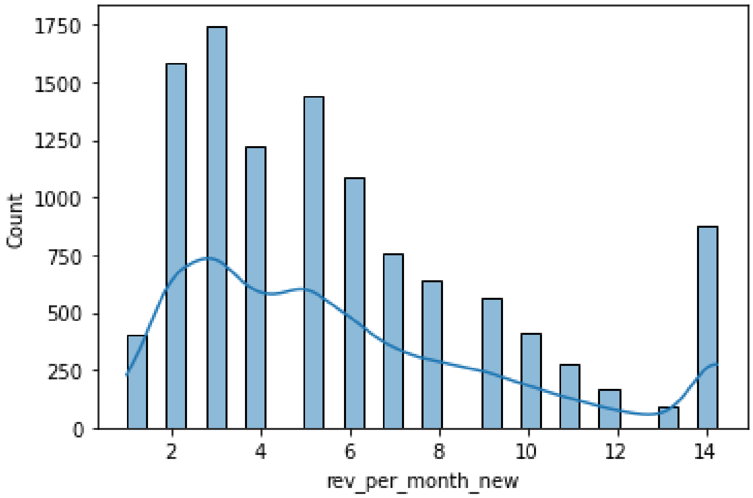


1a

In the above plot, it can be observed that there are very few outliers in rev_per_month; almost nill in rev_growth_you and there are more outliers in cashback where few are very much separated from other data points.

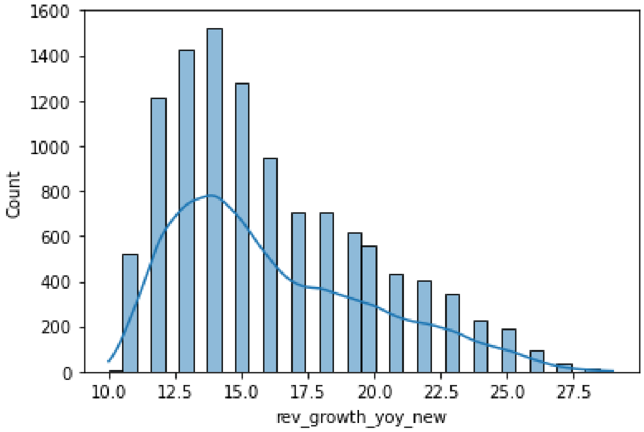
Treatment of outliers

Below is the distribution of rev_growth_month after outlier treatment



1b

Below is the distribution of rev_growth_yoy after outlier treatment

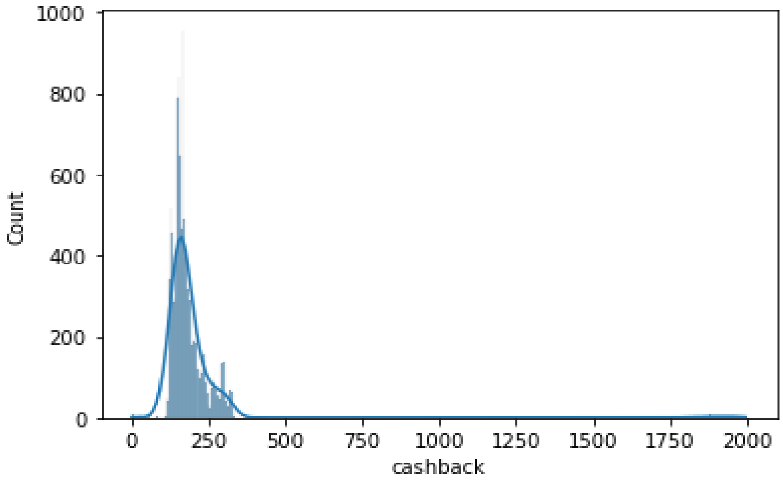


1c

There is no significant change in the distribution of variables, it can be said that outlier treatment has shown no result and original variable to be retained without treatment.

Treatment of outliers in cash back

Below is the distribution of cash back before outliers

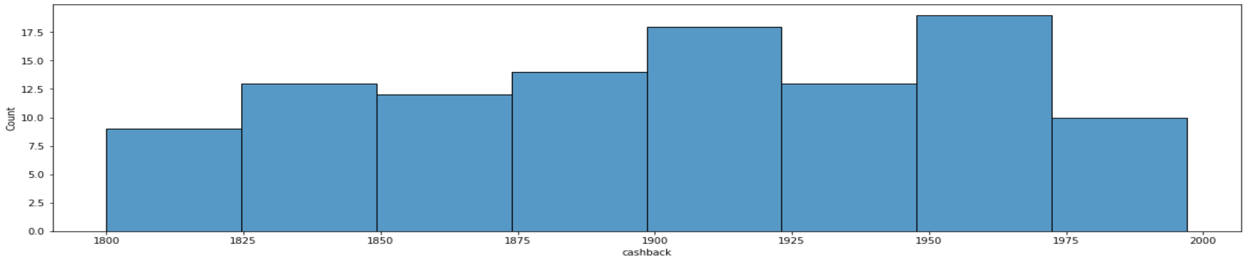


1d

The outliers in cash back are beyond a value 259, there are some data points which are accumulated or clustered beyond 1800 in the given data set.

These data points beyond 1800 are following a normal distribution when plotted.

Below is the distribution of points beyond 1800.



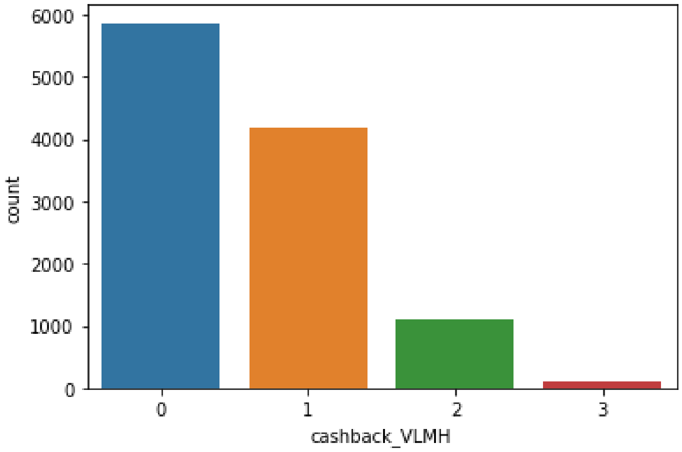
1e

From the above distribution plot, it can be observed that these values within them follow a normal distribution and do not have outliers within them.

It is also possible that these data points are collected from different source and merged in the data set, these are 107 in number, even though less in number there is possibility of different cluster formation and also can help model predict any such high cash back value in test data set.

Therefore, outlier treatment is not done on cash back value and is changed into ordinal categories 0 (<165.24), 1 (165.25 to 259), 2 (259 to 1800), 3 (>1800).

Below is the categorical cash back variable



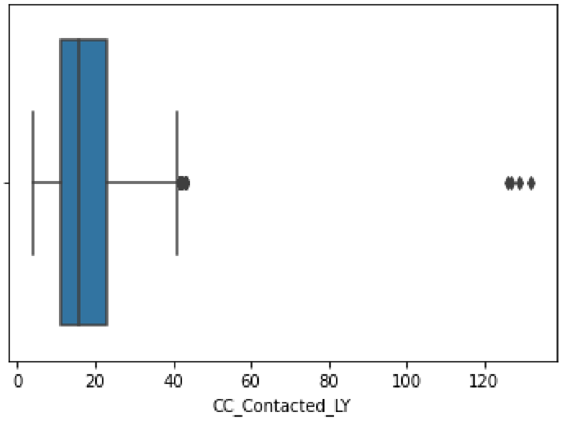
1f

The above is the cash back variable categorized based on the cash back customer received.

It can be observed that category 0 has high frequency followed by 1, 2, 3.

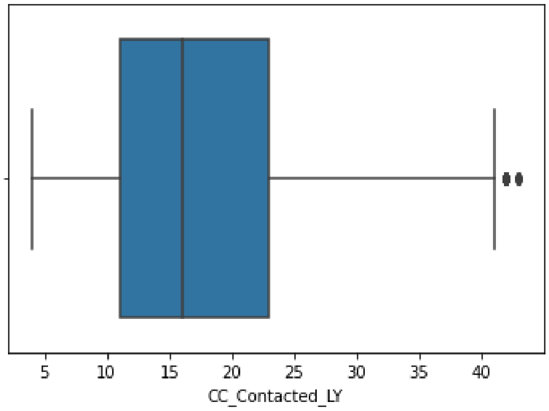
Outlier treatment in CC_Contacted_LY

Below is the boxplot of CC_Contacted_LY before treating outliers



1g

Below is the boxplot of CC_Contacted_LY after treating outliers



1h

For the above feature outlier treatment is done to improve distribution of data and ur,lr are not used to impute outliers, no changes have been made on few data points which are very close to upper whisker as they are not effecting the distribution of the feature.

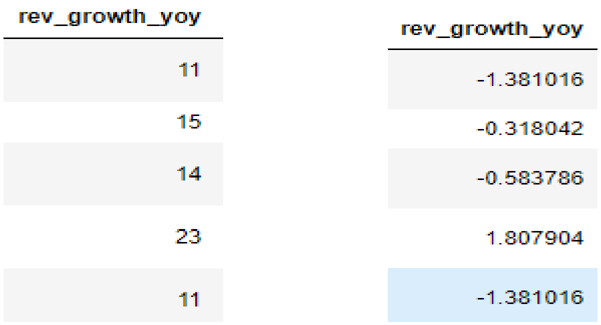
e) Variable transformation (if applicable)

Variable transformation is not required for the given dataset as most of the variables are categorical and there is no use in transforming categorical variables.

There are certain variables like rev_per_month , variable transformation on this variable will return negative values, it would add no meaning as a company cannot receive negative revenue.

Rev_growth_yoy is the only variable on which transformation can be applied and will be meaningful.

Rev_growth_yoy before transformation and after transformation

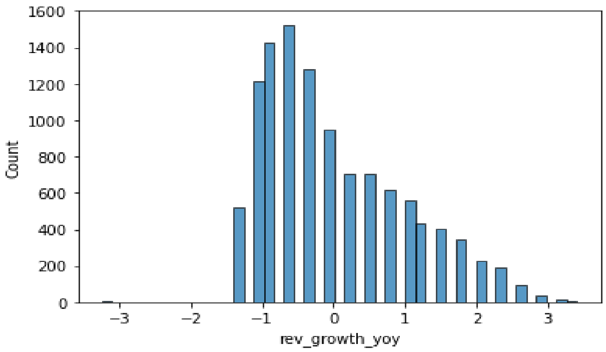


rev_growth_yoy
11
15
14
23
11

rev_growth_yoy
-1.381016
-0.318042
-0.583786
1.807904
-1.381016

1F

The left one is the variable before transformation and right one is variable after transformation.



1i

The above is the distribution of plot after transformation, where mean is 0.

f) Variables removed or added (if required)

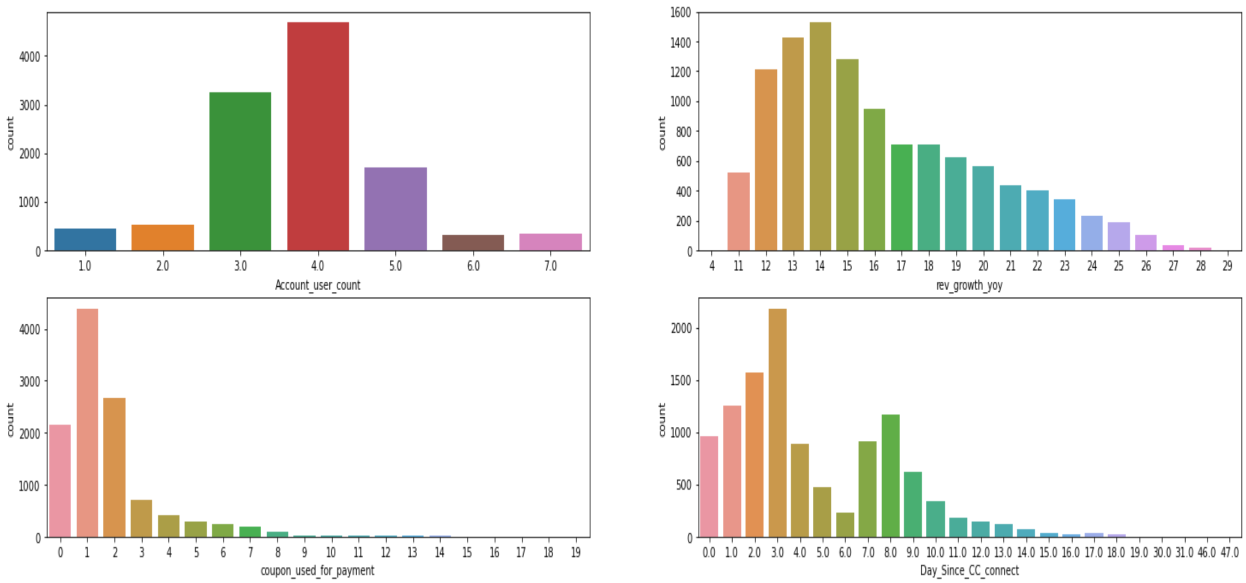
No new variables are added in the given problem statement.

Only change is categorization of cash back variable into ordinal.

****Note**: Below exploratory analysis has been done after data preprocessing to get best and unbiased insights from the data.

4. Exploratory Data Analysis

a) Univariate Analysis - Numerical data



1j

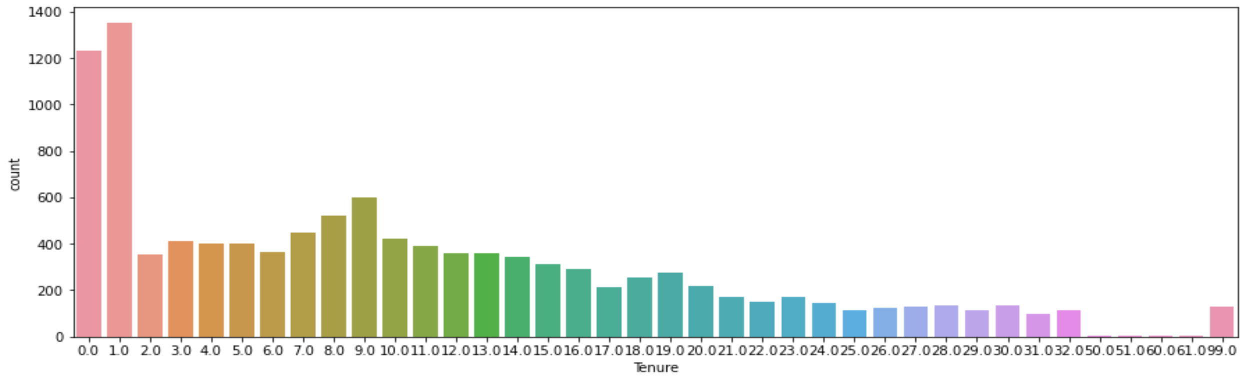
There are 4 customers connected to an account at maximum, followed by 3, 5.

Only one coupon is used for payment at maximum, followed by 2 coupons and most of the customers also have used no coupons for payment.

Revenue growth at maximum obtained is 14, followed by 13, 15, 12.

Maximum days since Customer connect is 3.0, which can be inferred that customers are contacting the customer care for every 3 days.

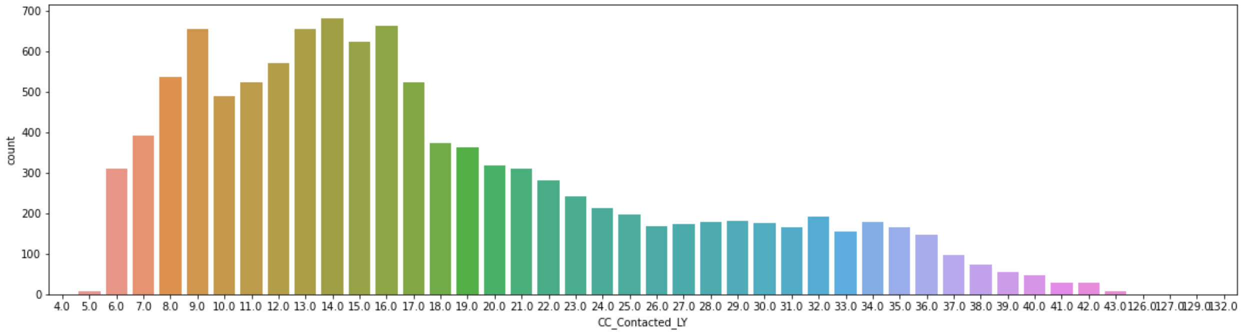
Tenure:



1k

The above is the count plot of Tenure, where it can be observed that there are more number of customers with 1.0, followed by 0 tenure and the frequency has reduced with increase in tenure. Most of the customers have maintained very less tenure.

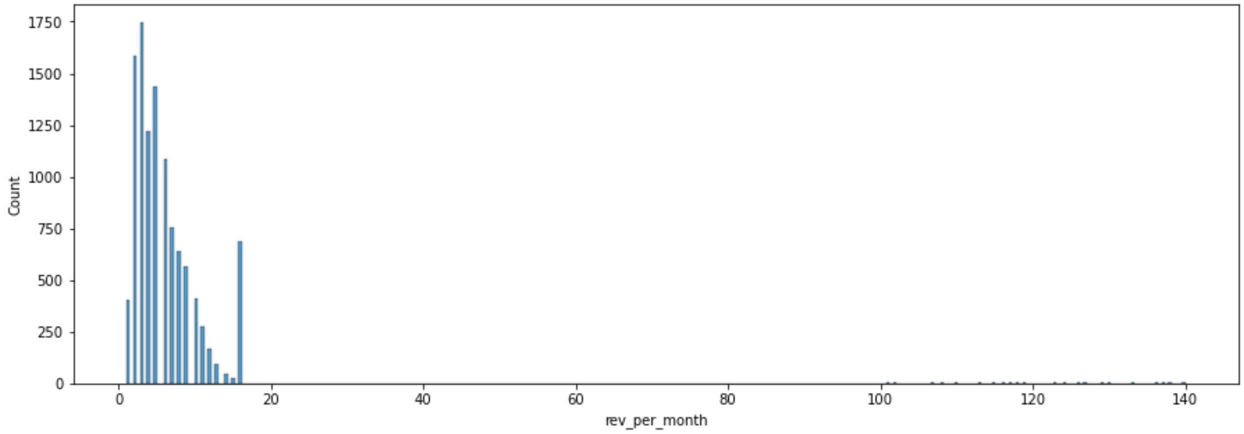
CC_contacted_LY



1l

From the above plot it can be observed that customers have contacted mostly between 8 to 17 times and have very less cases of contact beyond 40 times.

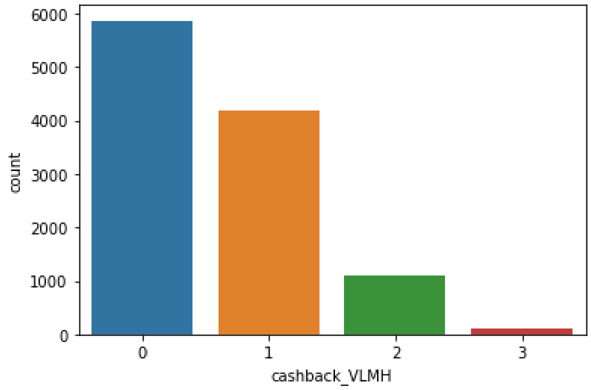
Rev per month



1m

The above is distribution plot of revenue per month, where it can be observed that average revenue is maximum in range of 0 to 10, the distribution is skewed to right side and showing decrease in frequency as the revenue is increasing.

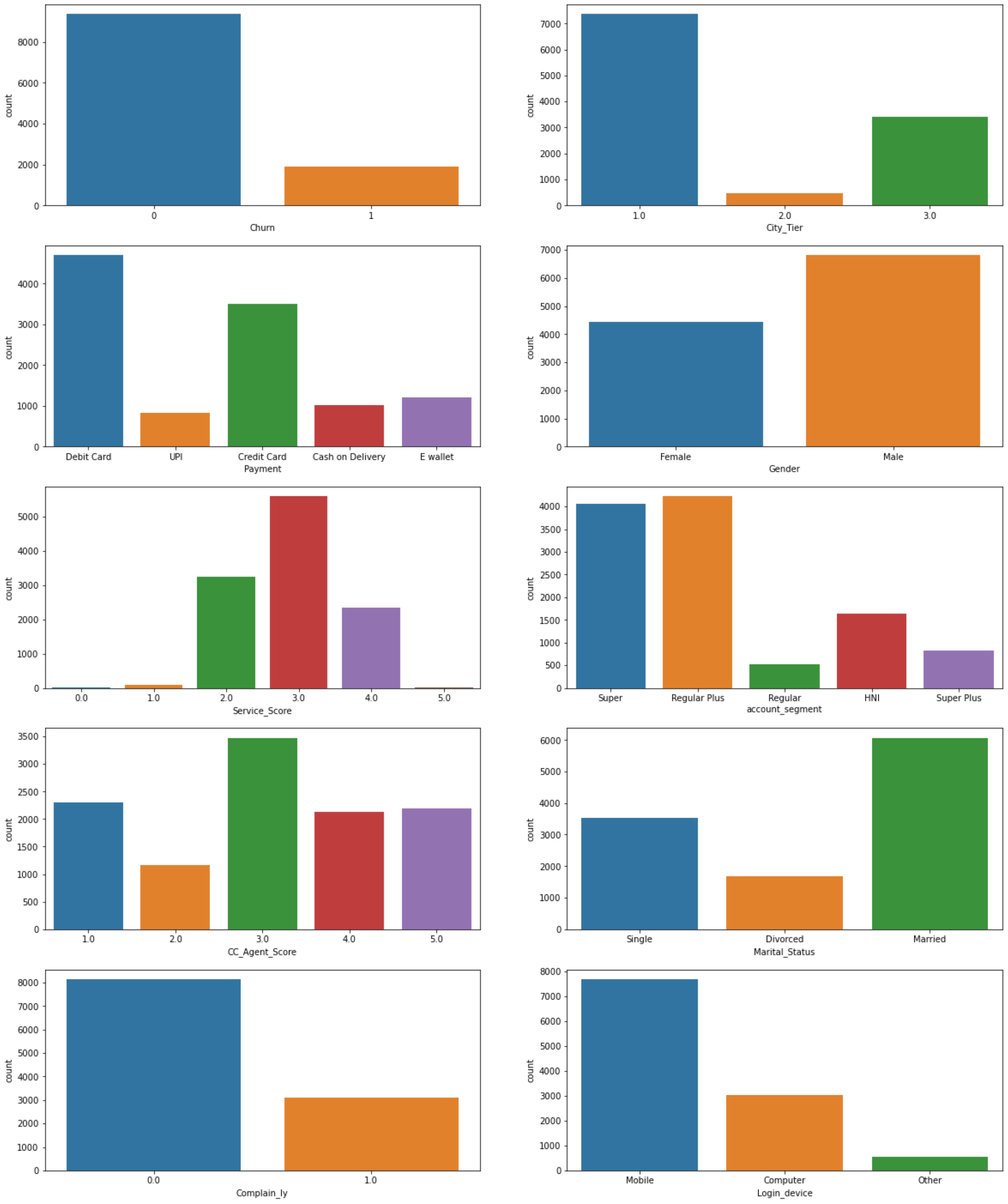
Cashback



1n

The above is cash back variable, where most of them are belonging to category 0, it means most of the customers are getting very less cash back, very few customers are getting high cash back.

b) Univariate Analysis – Categorical features



The above univariate analysis of categorical features.

Most of the customers are non-churners (83%), while very less are churners (17%).

Most of the customers come from Tier 1 city (64%) and very less from tier 2(4%).

It can be observed that payment mode preferred by most of the customers is Debit card(41%), followed by credit card (31%) and least chosen is UPI (7%).

Most of the customers are male (60%).

Most of the customers (73%) are from super and regular plus segment.

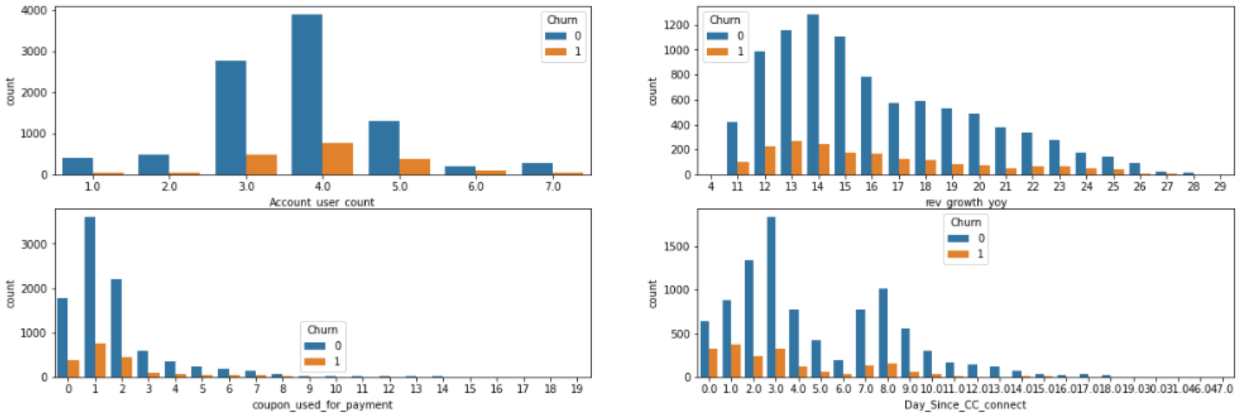
Most of the customers are married (53%), followed by unmarried(31%) and divorced(14%).

Most used device for login is mobile(68%), while other types of devices are very less used.

Most of the users (72%) have not raised complaints while very few(28%) have raised complaints.

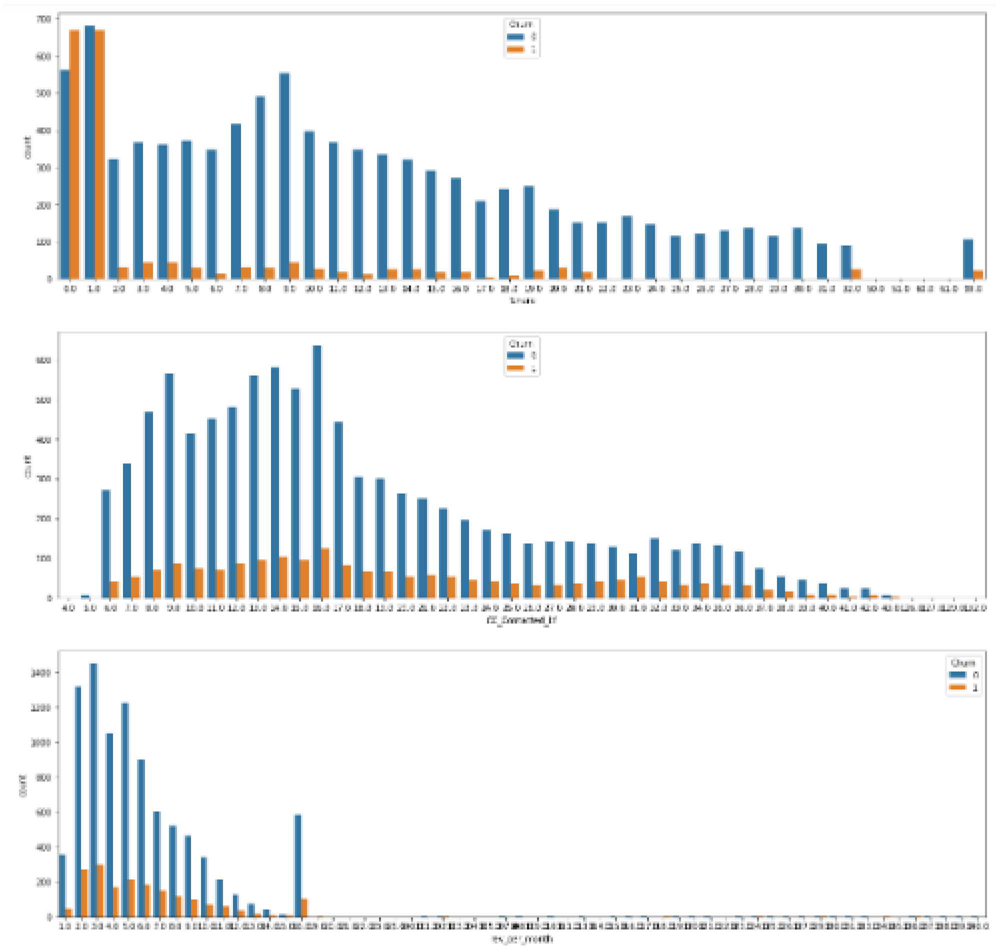
Agent score 3.0 (30%), 1.0 & 5.0 each at 20% , Service score 3.0 is most rated by customers.

c) Bivariate Analysis – Numerical variables



1p

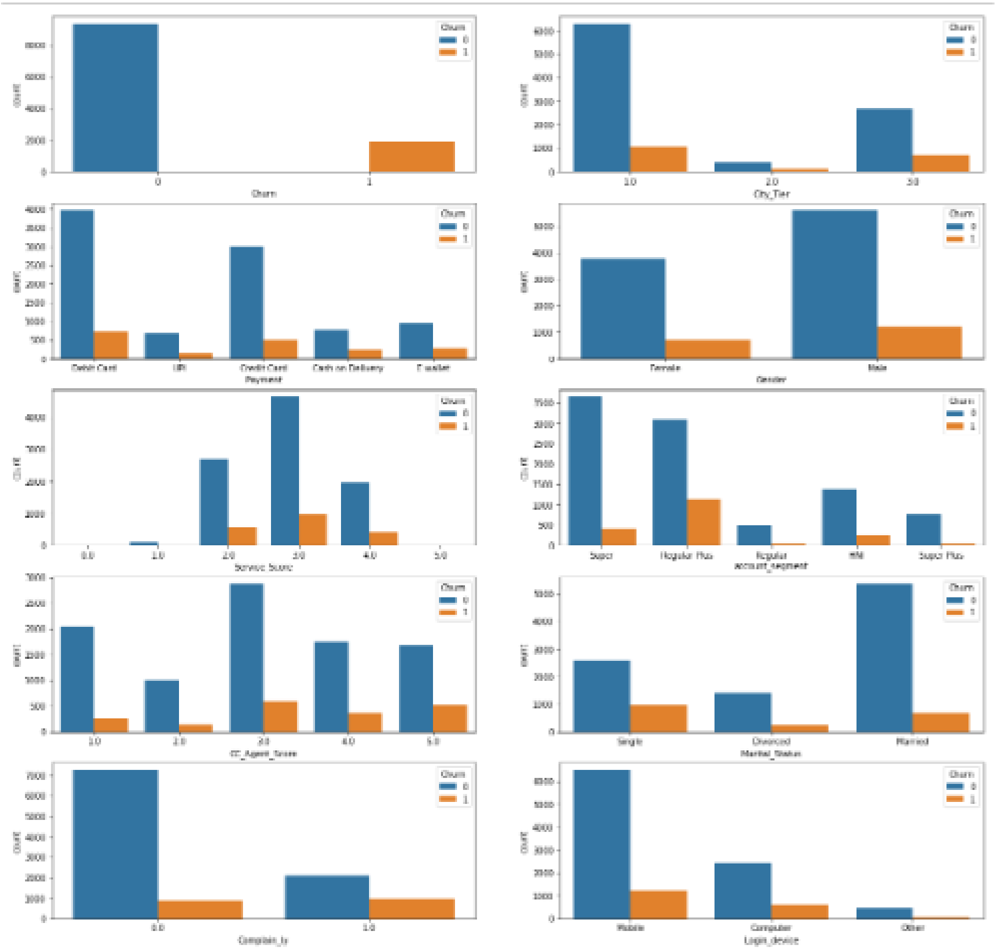
The above is bivariate analysis of few numerical variables, where it can be observed that all features have customers mostly not churned and very few customers have churned.



1q

From the above plot, there is very interesting trend observed that Tenure 0,1 have more churned customers, it infers that less the tenure more the chance to churn.

d) Bivariate – Categorical features



1r

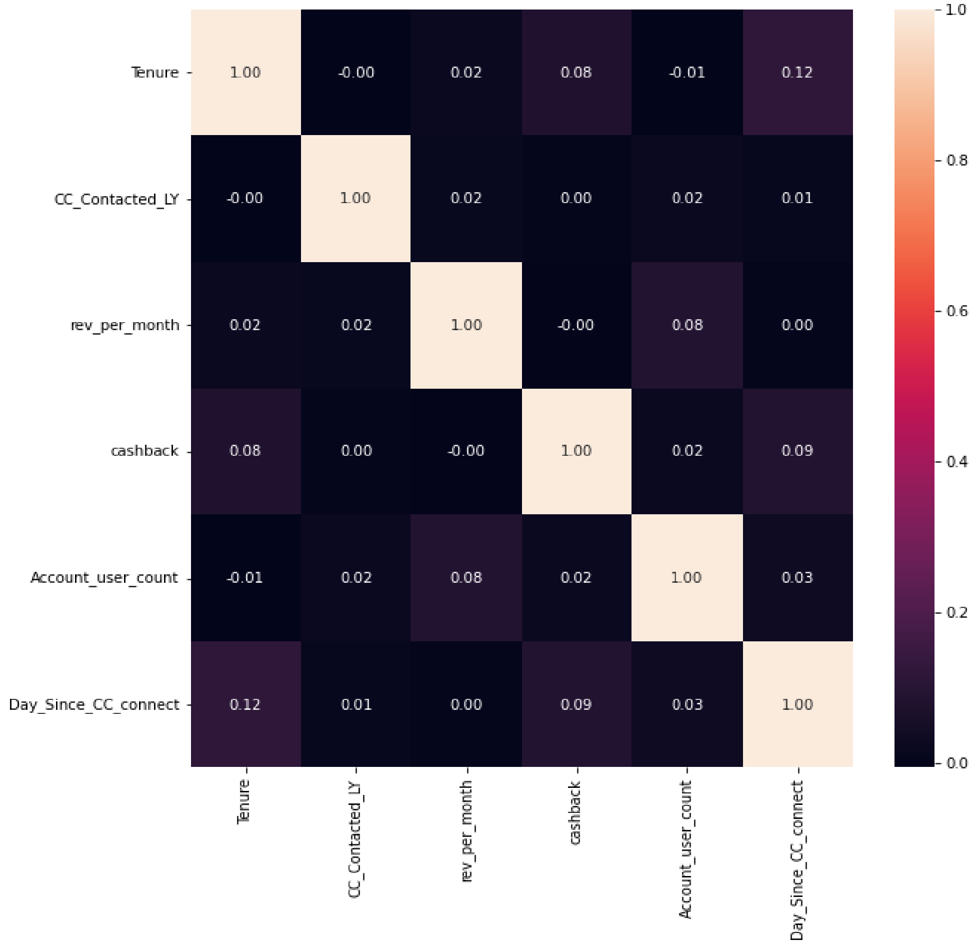
From the above plot it can be observed that all features mostly have customers who have not churned.

More number of customers have churned who have chosen regular plus account segment,

Churn rate is very less in customers who have not raised complaints.

Customers belonging to tier I have very less churn rate, Customers with Debit card, credit card as payment mode have less churn rate (<12%).

e). Multivariate Analysis



1s

Above is the heatmap of continuous variables, it can be observed that there is no Strong correlation between any two variables

c) Any other business insights

Most of the customers have not churned from the above analysis and most of them are from tier 1 cities, it can be observed that customers have contacted the customer care for every 3 days which have high chances to churn. Most of the customers are male, most used device is mobile, regular plus, super are most preferred and chosen segments.

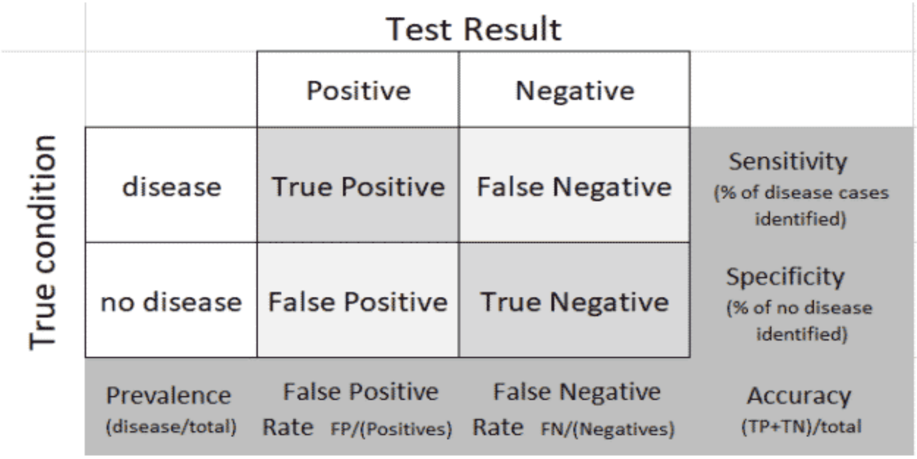
Less Tenure, more chances to churn.

5.Model building

Model Building Approach

The modelling approach used aims at predicting actual churners as churners for the given problem statement. In order to identify churned customers there is an important parameter called recall which needs to be considered in the confusion matrix of model.

Recall is also called as Sensitivity.



		Test Result		
		Positive	Negative	
True condition	disease	True Positive	False Negative	Sensitivity (% of disease cases identified)
	no disease	False Positive	True Negative	Specificity (% of no disease identified)
	Prevalence (disease/total)	False Positive Rate FP/(Positives)	False Negative Rate FN/(Negatives)	Accuracy (TP+TN)/total

1t

The above is confusion matrix which shows the actual, predicted positive values and negative values, recall is the True Positive (TP) predictions made by model out of actual values(TP+ FN).

Recall = True Positives / (True Positives + False Negatives)

Type I error = False Negative (Predicting churners as non churners).

For the given business problem, the model should perform with least Type I error that is minimum False negatives.

High the recall, less type I error.

There is also one more parameter called false positive rate which can be observed in ROC curve.

False Positive Rate (fpr) = FP/(FP+TN).

Less false positive rate, more true positive rate, high AUC score and there by customer churn prediction will be accurate.

F1 score is also an important parameter which reflects the balance between recall and specificity, high the f1 score better the model.

From the above model building approach it can be inferred that recall, f1 score, AUC score are important parameters which need to be observed in models.

*NOTE: SMOTE technique is not used to overcome the class imbalance, as it lead to over fitting.

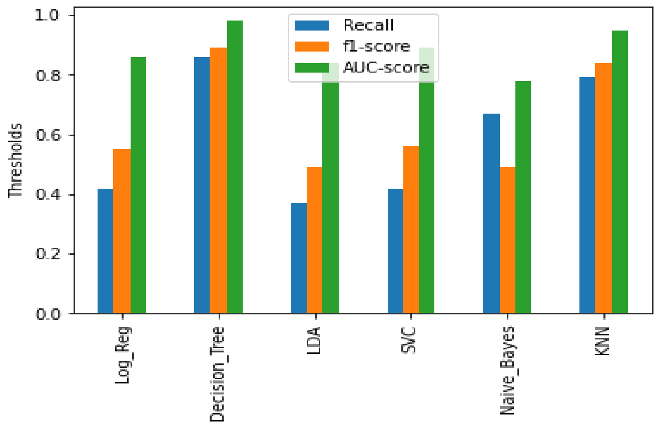
a.Build various models (You can choose to build models for either or all of descriptive, predictive or prescriptive purposes)

Below is the table containing parameters of all models in train data

Train Data					
Model	Accuracy	Precision	Recall	f1score	AUC score
Logistic Regression	0.88	0.78	0.42	0.55	0.86
Stats Model	0.84	0.55	0.24	0.34	-
Decision Tree	0.96	0.92	0.86	0.89	0.98
Linear Discriminant Analysis	0.87	0.73	0.37	0.49	0.84
Support Vector Classifier	0.89	0.84	0.42	0.56	0.89
Naive Bayes	0.77	0.39	0.67	0.49	0.78
KNN	0.95	0.89	0.79	0.84	0.95

1G

Comparison of various models parameters on train data



1u

From the above table and plot it can be observed that decision tree has performed very well with high values in recall, f1 score, AUC score in train data, the models performance should be tested on test data.

6. Model validation

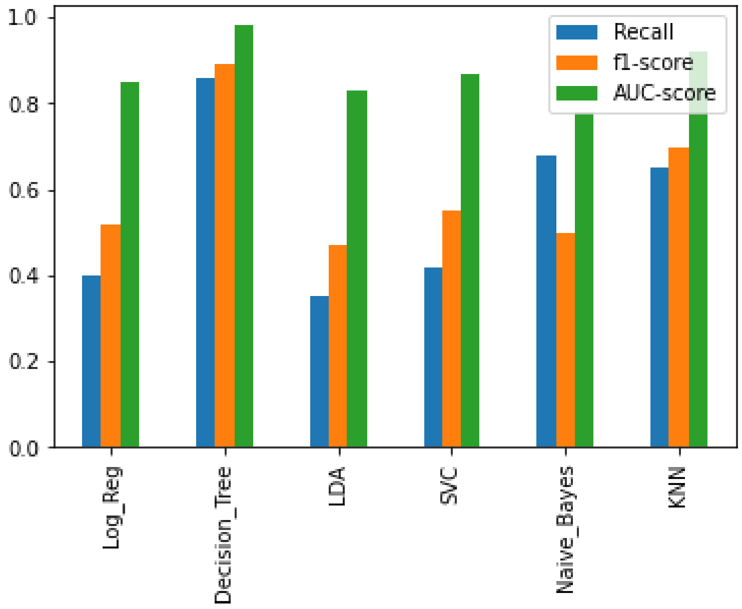
a.Test your predictive model against the test set using various appropriate performance metrics

Below is the table containing parameters of all models in test data

			Test Data		
Model	Accuracy	Precision	Recall	f1score	AUC score
Logistic Regression	0.88	0.77	0.40	0.52	0.85
Stats Model	0.84	0.60	0.24	0.34	-
Decision Tree	0.96	0.92	0.86	0.89	0.98
Linear Discriminant Analysis	0.86	0.72	0.35	0.47	0.83
Support Vector Classifier	0.88	0.82	0.42	0.55	0.87
Naive Bayes	0.77	0.40	0.68	0.50	0.78
KNN	0.90	0.75	0.65	0.70	0.92

1H

Comparison of various models parameters on test data



1v

Comparision of models:

As discussed above parameters used for validation are Recall, f1score, AUC score

All models have performed very well in terms of accuracy, both train and test accuracies are high for every model.

Only Decision tree, KNN models were able to balance between precision and recall reflected by high f1 score, while all others model specifically LDA has least f1 score.

Recall is high for Decision Tree, Naïve Bayes and KNN. These three models have predicted more true positives (actual churners as churners) compared to other models.

There is no big difference between recall & f1 scores of Decision Tree & KNN, while huge difference is identified in Naïve Bayes.

When both train and test parameters are observed for all models, there are no over fitting and under fitting observed, all models have balanced bias variance trade off.

Out of all models, Decision Tree and KNN have performed well in terms of all keys parameters like recall, f1 score, AUC score.

Conclusion

From the above plot it can be observed that decision has performed well on test data also.

It has achieved highest recall, fi score, AUC score compared to all models . The decision has predicted 86% of actual churners as churners, it has 89% of f1 score reflecting good balance between precision & recall, It also has very high AUC score 98% inferring less false positive rate, high true positive rate and model has made very less false predictions.

b.Interpretation of the model(s)

***NOTE: All models are tested at different thresholds but none of them improved performance of model, threshold value 0.4 is considered for stats model and 0.5 default is considered for other models.**

Interpretation of Logistic Regression :

Logistic Regression model has achieved accuracy of 88% in both train and test sets, AUC sc ore (Area under curve) is 85%, which infers that the model has predicted true and false pre dictions upto 85%, but the model couldn't balance well between precision and recall reflect ed in f1 score. The positive predictions out of total actual predictions is very poor which is only 24%. It can be said that Logistic Regression failed to resolve multiclass or multi catego ry problem and couldn't separate the classes 1 Vs 0 well.

Why Logistic Regression?:

Logistic Regression is used as the target to predict is categorical; it uses sigmoid function to separate data points to classes at an optimum threshold.

Interpretation of Stats Model :

Stats model is one of the best classification model, which being white box models explains the inner logic of selection of features and effect of features on dependent variable. There is no improvement in performance of model even after using significant features in model building. All parameters like precision, recall f1 score are minimum compared to all models, it can be said the logit or sigmoid function failed in separating multiple classes.

Why Stats Model:

Stats model is considered to understand the effect of features on dependent variable and h ow much they are significant in predicting target. It being white box model explains the sig nificance of feature through hypothesis.

Interpretation of Decision Tree :

Decision Tree has acheived highest accuracy 96% on both training and test sets, not only the model has predicted true positives (1 as 1) and false negatives correctly(0 as 0), it also had predicted the false predictions accurately with almost 0 false positive rate at threshold 1.0, which is reflected by 98% AUC score, the model has minimized scope of type I and type II errors. Decision Tree used gini purity for classification criterion and has in fact classified with highest accuracy compared to all above models.

Why a Decision Tree:

Decision Tree is chosen as one of the models for the given problem because, the target to predict is classification type and decision tree with proper hyper parameters and pruning performs classification well with gini impurity as criteria.

Decision Tree has also resulted in high recall which is expected for the given problem. It has also balanced well between bias and variance which can be observed from high accuracy in test set 96%.

Interpretation of Linear Discriminant Analysis:

LDA has achieved high accuracy in both train and test sets and false positive rate was also moderate in the model but not 0 at threshold 1.0, parameters like recall, precision, f1 score are minimum compared to all models. It can be observed that LDA failed to capture and minimize variances between two classes means in multi-dimensional space and thereby couldn't separate the classes well, this might happen as almost all features in the data are categorical.

Why LDA :

LDA is known as classification type supervised machine learning algorithm, it separates classes in different dimensional spaces.

Interpretation of Support Vector Classifier:

Support Vector model has achieved good accuracy 89%, and also has less false positive rate with AUC score 89%, the model has made high type I error that is predicting churn as not churn. The model couldn't balance well between recall and precision which can be reflected in f1 score(0.56). Support Vector uses hyper plane with margins to separate classes, wider the margins, high the confidence, even after using kernel function radial basis function, the model couldn't predict to the best, the multiclass problem couldn't be resolved by SVC.

Why SVM:

SVM is used in case of classification problems and is best suited for the given problem statement, it separates classes using hyper plane or decision boundary, high the margin width, better the classes separation.

Interpretation of Naive Bayes:

Naive Bayes classifier works very well for classification problems, but imagines all features are independent to each other which is not possible in large data sets, the model didn't achieve high accuracy but has very good recall compared to precision, it infers that model has predicted 67% of actual churn as churn. AUC score is least compared to all models infers that the false positive rate is high and model couldn't predict false predictions correctly. It can be said that Naive Bayes has least performed out of all models.

Why Naïve Bayes:

Naïve Bayes is used for classification problems which uses prior probability to identify posterior probability. It is applicable for the given business problem as target feature and most of the independent features are categorical.

Interpretation of KNN:

From Both train and test data, it can be observed that the model is over fitting in train data and least performed on test data, still the model has performed very well with accuracy 90% and AUC score 92%. The model has balanced well between Recall and precision with f1 score 0.70. It can be said KNN has predicted well but not to the best accuracy.

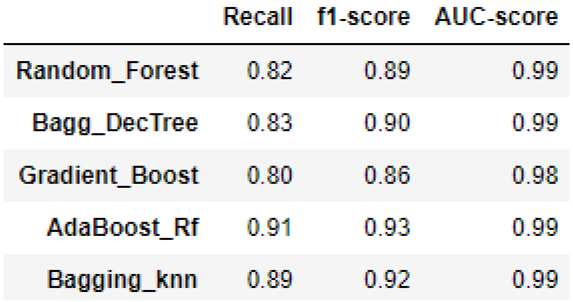
From the above interpretations of all models it can be observed that Decision Tree is the best model followed by KNN.

Why KNN:

KNN is used to classify a data point basing on n nearest neighbors of the data point, it is applicable for the given business problem, as it is classification type.

c.Model Tuning - Ensemble modeling wherever applicable

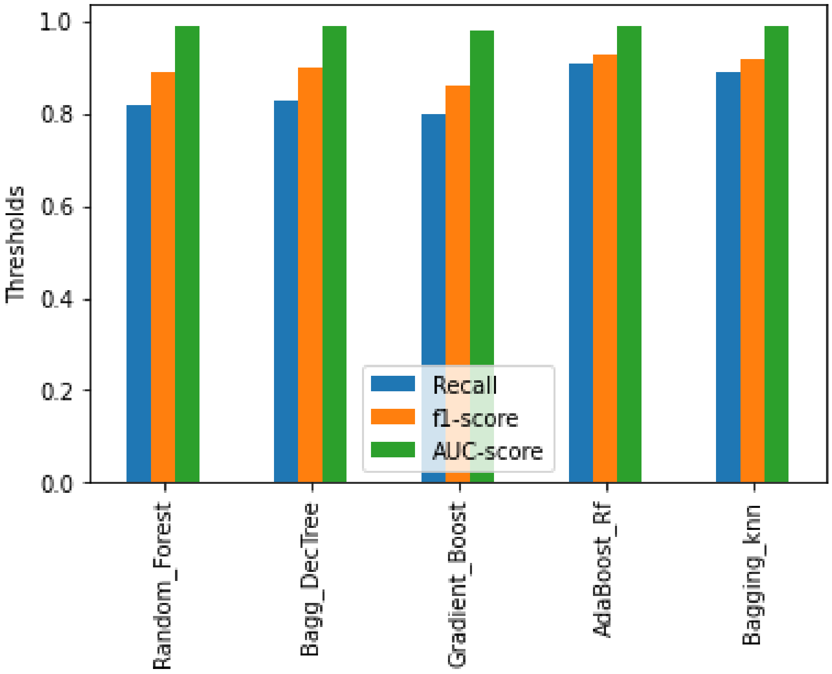
Below are few ensemble models built using train data



	Recall	f1-score	AUC-score
Random_Forest	0.82	0.89	0.99
Bagg_DecTree	0.83	0.90	0.99
Gradient_Boost	0.80	0.86	0.98
AdaBoost_Rf	0.91	0.93	0.99
Bagging_knn	0.89	0.92	0.99

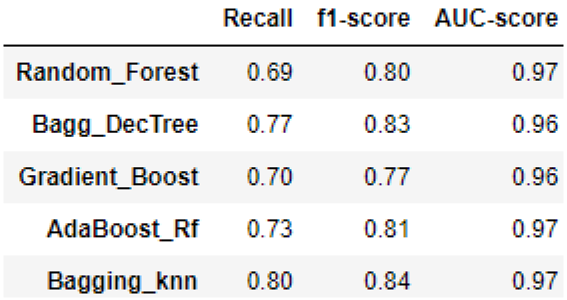
1I

Comparison of various models parameters on train data



1w

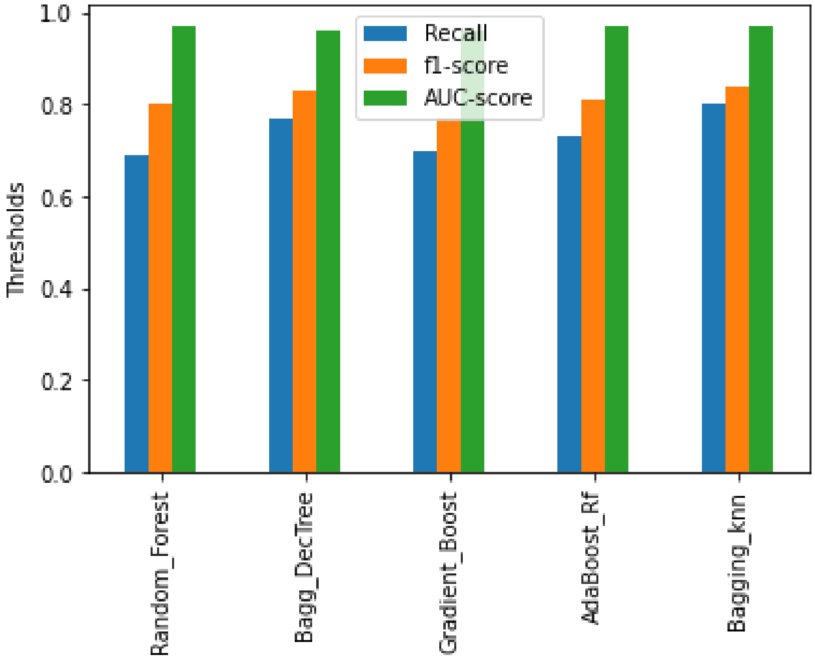
Below are few ensemble models built using test data



	Recall	f1-score	AUC-score
Random_Forest	0.69	0.80	0.97
Bagg_DecTree	0.77	0.83	0.96
Gradient_Boost	0.70	0.77	0.96
AdaBoost_Rf	0.73	0.81	0.97
Bagging_knn	0.80	0.84	0.97

1J

Comparison of various models parameters on test data



1x

d. Measures to improve model performance

Above are few ensemble models used as an attempt to improve model performance of best models like decision tree, KNN .

Bagging with decision tree and random forest as base estimators are used with various model tuning parameters like max depth, random state, number of estimators, min samples split, min samples leaf.

As maximum depth increased there is over fitting observed, models performed well on train data and couldn't perform well on test data, the most effected parameters due to over fitting are recall, f1 score.

As given problem statement has mostly categorical data, entropy which includes logarithms has shown minimal effect in performance, while gini purity is better choice as criterion for better classification of nodes in both Decision Tree and Random Forest.

Out of bag score is also important metric , which is True for few models, where no entries are missed in any of the models in training. Even though this parameter is used in ensemble modeling, it couldn't improve model performance.

Considering KNN, minkowski is the best metric and default five nearest neighbors is recommended. As it is classification problem, odd number of neighbors ensures accuracy in majority voting.

Ada Boosting, Gradient Boosting and Random Forest techniques have less performed due to over fitting, the models failed to perform well on test data. These models were trained with certain classification in train data which they couldn't find in the test data.

Conclusion:

From the above analysis it can be observed that efforts made to improve model performance using ensemble techniques are not fruitful. Neither Decision Tree nor KNN performed well in ensemble models.

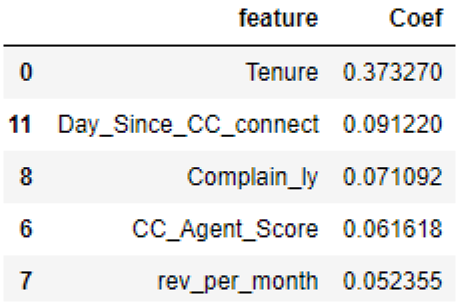
7. Final interpretation, Analysis, Insights and recommendations.

Interpretation of the most optimum model .

In order to select best model, various parameters like Recall, f1score, AUC score are to be considered. As discussed above in the model building approach, the model with high recall, f1score and AUC score is considered as best model, the best model makes less type I error, less false predictions and thereby can predict churned customers exactly as churners.

Out of all the models Decision Tree has performed well, it has not only predicted churners well but also balanced bias variance trade off with no over fitting and under fitting. Hyper parameters used in building the model have pruned the decision tree to proper depth and resulted in the best model.

Important Features

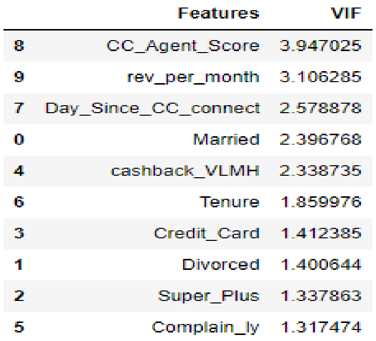


	feature	Coef
0	Tenure	0.373270
11	Day_Since_CC_connect	0.091220
8	Complain_ly	0.071092
6	CC_Agent_Score	0.061618
7	rev_per_month	0.052355

1K

The above are the top 5 features identified by the decision tree model as most significant in deciding the dependent churn.

Multi Collinearity within features explained by VIF (Variance Inflation factor).



	Features	VIF
8	CC_Agent_Score	3.947025
9	rev_per_month	3.106285
7	Day_Since_CC_connect	2.578878
0	Married	2.396768
4	cashback_VLMH	2.338735
6	Tenure	1.859976
3	Credit_Card	1.412385
1	Divorced	1.400644
2	Super_Plus	1.337863
5	Complain_ly	1.317474

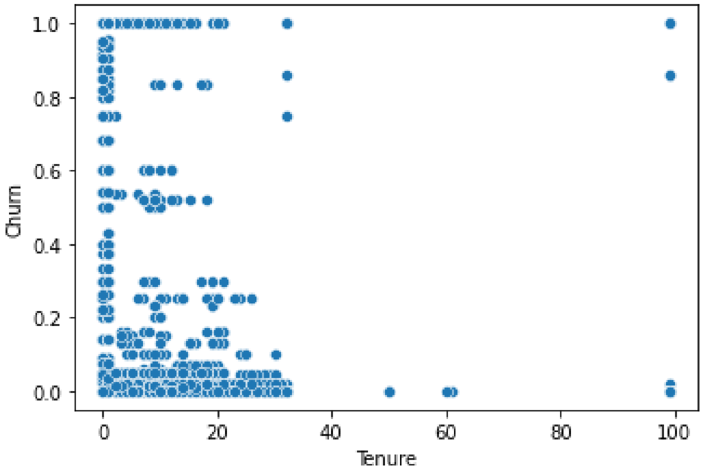
1L

VIF infers multi collinearity of a feature with other features, it can be observed that there is no high value in VIF (>5) and no feature is highly collinear with other features.

It can be inferred that coefficients in the important features table are not inflated and individually effect the target.

Analysis of features after prediction

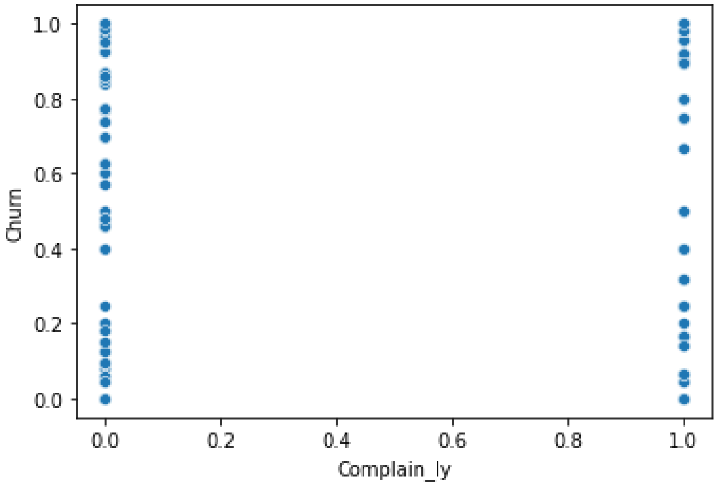
Tenure Vs Churn :



1y

From the above plot, it can be observed that on y axis all data points below 0.5 are more likely to not churn (0) and above 0.5 are likely to churn (1). For range of tenure between 0 & 40, more customers are not likely to churn. As tenure increases churn rate reduces.

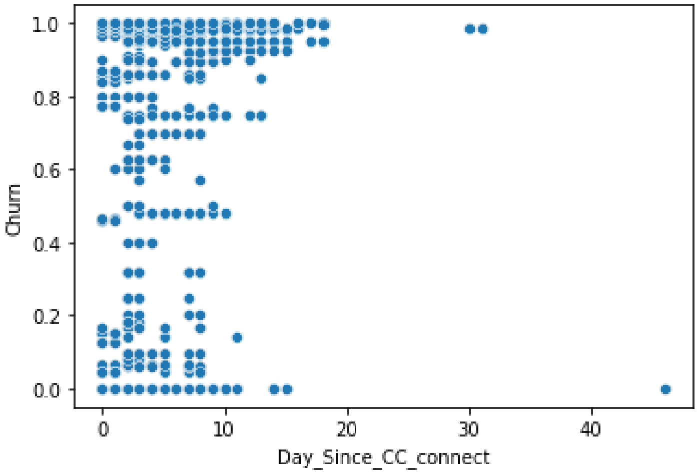
Complain_ly Vs Churn :



1z

From the above plot it can be observed that both complaints raised and not raised have equal chances to churn and not churn.

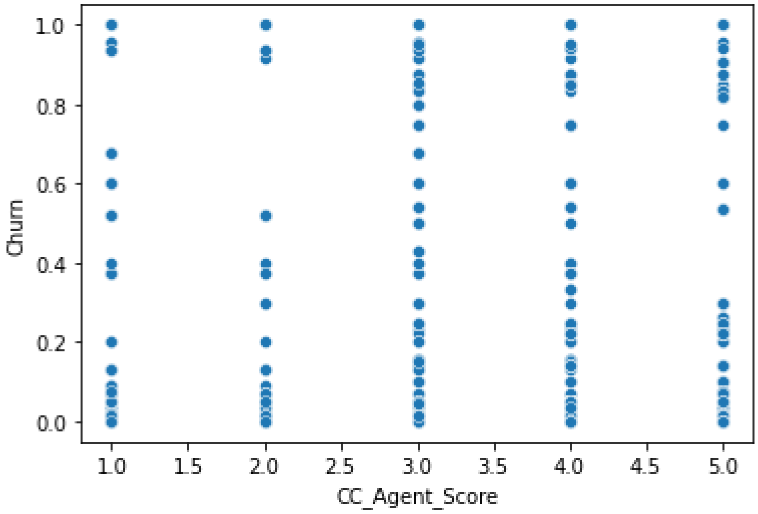
Day Since Customer Connect Vs Churn:



2a

From the above plot it can be observed that for same range of days since customer connect there are more chances to churn when threshold 0.5 is observed on y axis. It can be inferred that on frequent contact to customer connect, there are more chances to churn.

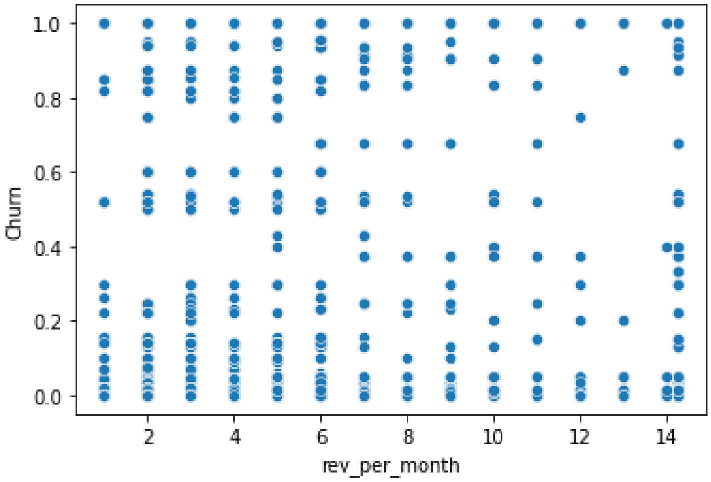
cc Agent score Vs Churn:



2b

From the above plot it can be observed that low Agent score has resulted in less chances to churn and Agent score at 3,4 have high chances to churn.

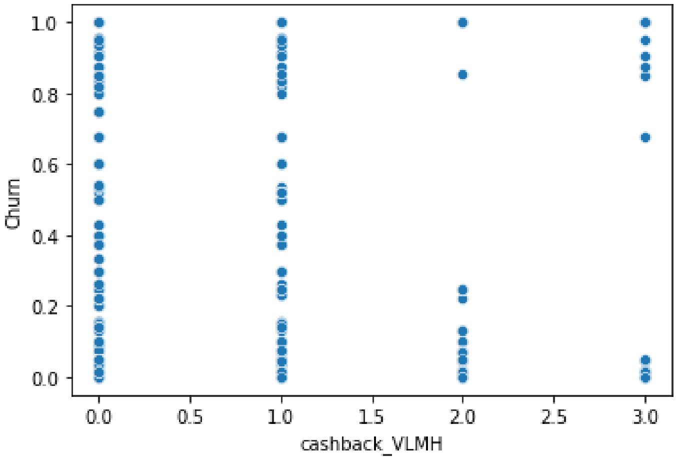
Revenue per month Vs Churn:



2c

From the above plot it can be observed that there are equal chances to churn and not churn till revenue of 8 and beyond it there is less chances to churn, it can be inferred that more revenue per month and less chances to churn.

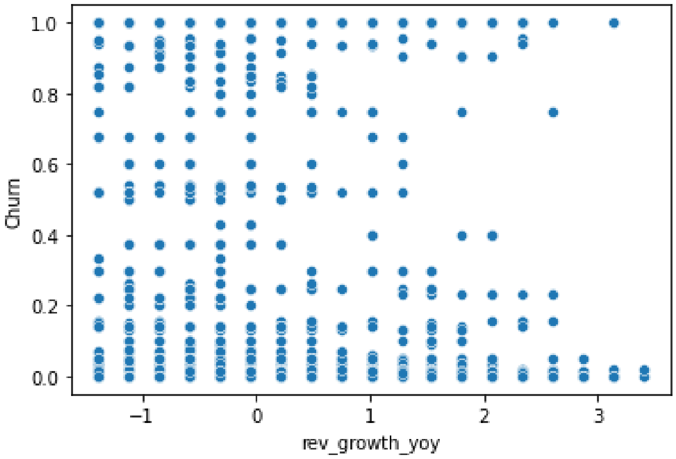
Cashback_VLMH Vs Churn:



2d

From the above case it can be observed that cash back Very Low (0), Low(1) have equal chances to churn, while medium cash back and high cash back to account have less chances to Churn. It can be inferred that high the cash back amount for customers, less chances to churn.

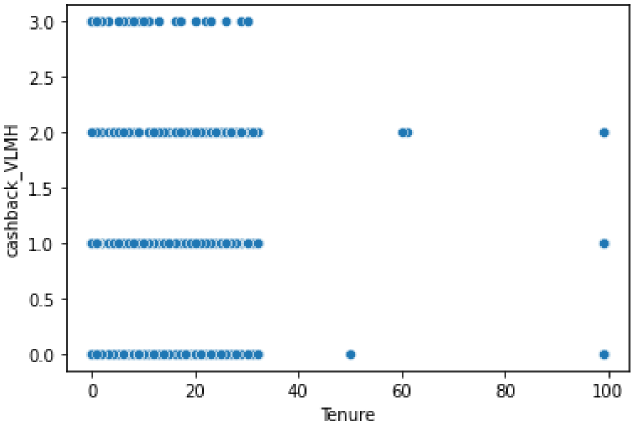
Revenue growth yoy Vs Churn:



2e

It can be observed that there are equal chances to churn and not churn at all thresholds including 0.5. As the revenue growth increased , there is decrease in chances of churn and also high confidence in classes separation observed.

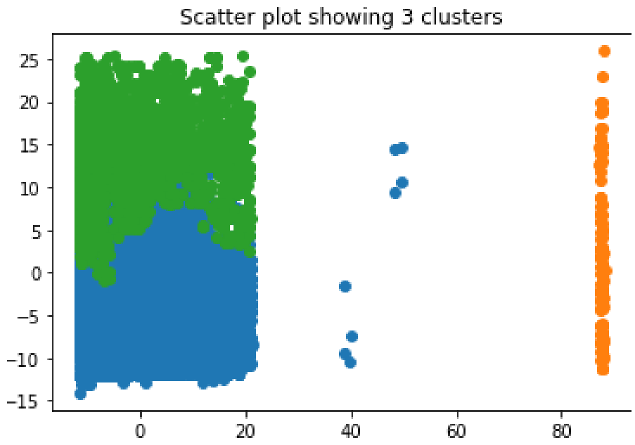
Cashback_VLMH Vs Tenure:



2f

More Tenure has resulted in low cash back in few cases, while in most of the cases the cash back is same in all levels in Tenure ranging between 0 & 40.

Cluster Segmentation of customers:

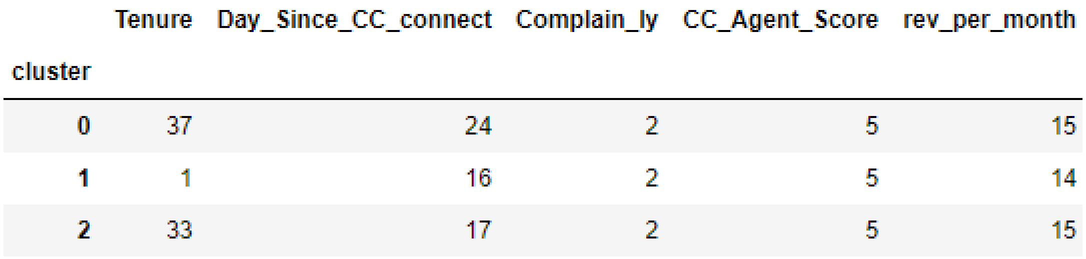


2g

The above plot displays the segmentation of customers into 3 groups. It can be inferred that few data points in blues cluster are far away from the other group.

Orange cluster is quick different from other two clusters.

Important Features in each Cluster



cluster	Tenure	Day_Since_CC_connect	Complain_ly	CC_Agent_Score	rev_per_month
0	37	24	2	5	15
1	1	16	2	5	14
2	33	17	2	5	15

1M

From the above table it can be inferred that customers in Cluster 0 have more Tenure, have more days since cc connect. This cluster has high chances of low churn rate.

Cluster 1 has very less Tenure and very less number of days since cc connect on average, revenue per month is also low compared to other clusters. Customers churn possibility is high.

Cluster 2 also have more Tenure after Cluster 0 and days since cc connect is better than Cluster 1.

It can be said that customers from Cluster 0 are best customers, followed by Cluster 1, 2.

Insights from Analysis

1. From the above analysis it can be inferred that all features have less VIF (<5) and there is no inflation in coefficients due to multi-collinearity.

2. Most Significant features that effect the churn are Tenure, Complain_ly, Day since cc connect, revenue per month, Cc agent score, cashback_VLMH.

3. Customer accounts with more tenure and more number of days since customer agent connect have minimum chances to churn. It can be inferred that customers who are using the services for long time have very less concerns, queries and also are interested in services offered by companies.

4. Customers with more tenure have less cash backs, it can be observed that customer on long relation with company are least expecting any special cash backs, offers and continuing with existing plans.

5. Less number of Days since cc connect infers customer dissatisfaction with services and more chances to churn.

6. Customer with High and Medium category cash back values have least chances to churn.

7. More revenue per month made many data points fall below the threshold 0.5 and there by less chances to churn.

8. Less Agent score has leaded to less chances of churn, this might be due to issue escalation and quick resolution of query involving high officials satisfied the customer and avoided customer churn.

Recommendations

1. The company needs to focus on increasing the tenure of the customer, reducing number of tickets raised, immediate ticket resolution, these will improves confidence on services, helps to extend relation with the company and there by reduces churn rate.

2. Customers who have received high cash back/ offers have less chance to churn, keeping in mind revenue generated; the company can focus on giving offers to retain existing customers.

3. Company can explain DTH plans to customers including the service charges, which not utilized can be offered as cash back into customer account for certain time and later when not utilized in long term can be considered as profit.

4. The profits generated above can also be invested in creating new customers and improving services. This will impose minimum stress on revenue assurance team to implement new plans for new customers.

5. Company has maintained consistency in revenue growth year on year, but can also focus on increasing revenue.

6. Even though Cluster 0 has more Tenure, days since cc connect, the revenue groth of it is same as Cluster 2, company can focus to increase number of customers in Cluster 0 by providing better service.

7. Customers from Cluster 1 have more chances to churn, it is recommended for the company to offer more cash back and continuous assistance by customer support.

8. The data provided has no diversity within itself which infers that most of the customers are using similar plans or come from similar demography. Company needs to implement different plans according to needs of customers according to demographics and analyse growth trends, customer experience.

9. Considering the revenue, company can provide free subscriptions for sports channels especially during world cup seasons, this can be temporary till it can retain customer and can remove once the customer account activity is consistent and tenure is increased.