Telecom Churn Case Study

Aswin Kumar. K

Anupriya Kumar

Anuradha Subramanian

Problem Statement

- Telecom industry is a highly competitive market, which experiences an average of 15-25% annual churn rate
- Retention of high profitable customers is the number one business goal
- **Objective:** ***To predict the churn in the last month using the data (features) from the first three months***
- Critical data points to consider:
- The ‘good’ phase: In this phase, the customer is happy with the service and behaves as usual. (First 2 months)
- The ‘action’ phase: The customer experience starts to turn sore in this phase (Third month)
- The ‘churn’ phase: In this phase, the customer is said to have churned (fourth month)

Methodology used for the study

01. Data cleaning and preparation

Handling missing values

Mapping categorical variables to integers

02. Test-train split and Scaling

Data split into 70 to 30 ratio

SMOTE used for class imbalance

03. Model Building

Logistic regression with PCA

Tuning hyperparameters

04. Model Evaluation

Random forests with PCA

Optimal cutoff using ROC

Precision and Recall

05. Prediction on test set

Final model testing for results

01. Data cleaning and preparation

Total data – 226 columns and 99999 rows

Only 12 columns/categories are object type

Data cleaning key steps

Columns with high missing values; 70% is considered the threshold.

low significance columns are dropped – ID and date columns

Dropped values with 1 unique NaN

Zero values imputed in the columns of missing values in recharge columns

Data preparation

Creating column avg_recharge_6_7 by adding total recharge amount of 6 & 7 month, then take avg of sum

Tagging the CHURNERS

Removed outliers less than 10th and more than 90th percentile

01. Data cleaning and preparation

Exploratory Data analysis
data visualizations

Adding New Columns with insights for predicting churn

dec_rech_action: Indicating if the number recharges for a customer when compared to good phase has decreased in action phase or not

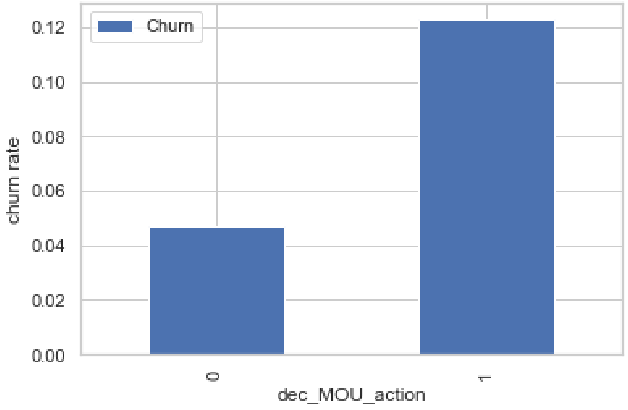
dec_avg_revenuePC_action: Indicating if average revenue per customer when compared to good phase has decreased in action phase or not

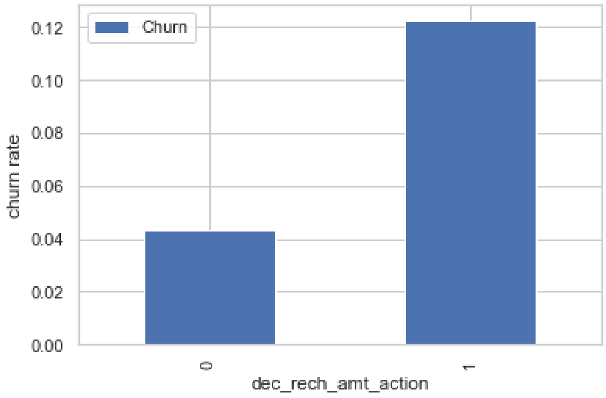
dec_rech_amt_action: Indicating if recharge amount of customers when compared to good phase has decreased in action phase or not

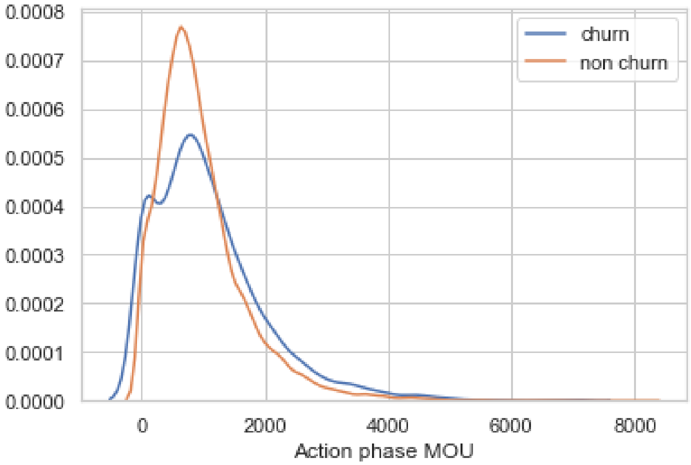
dec_MOU_action: Indicating if minutes of usage of customers when compared to good phase has decreased in action phase or not

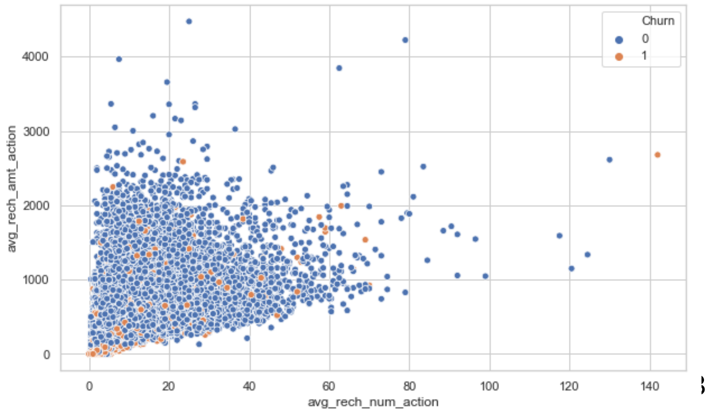
01. Data cleaning and preparation

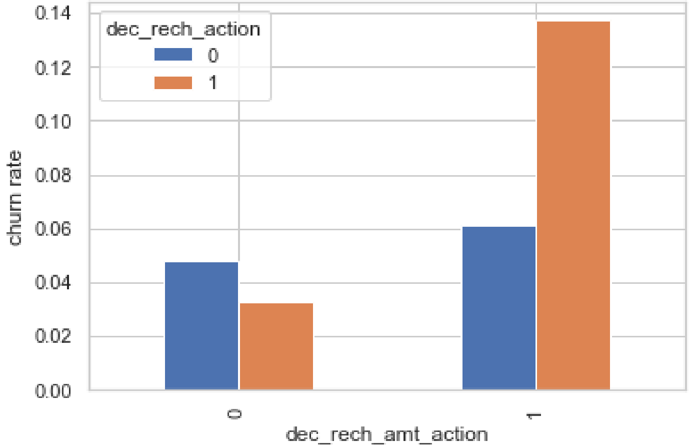
Exploratory Data analysis data visualizations





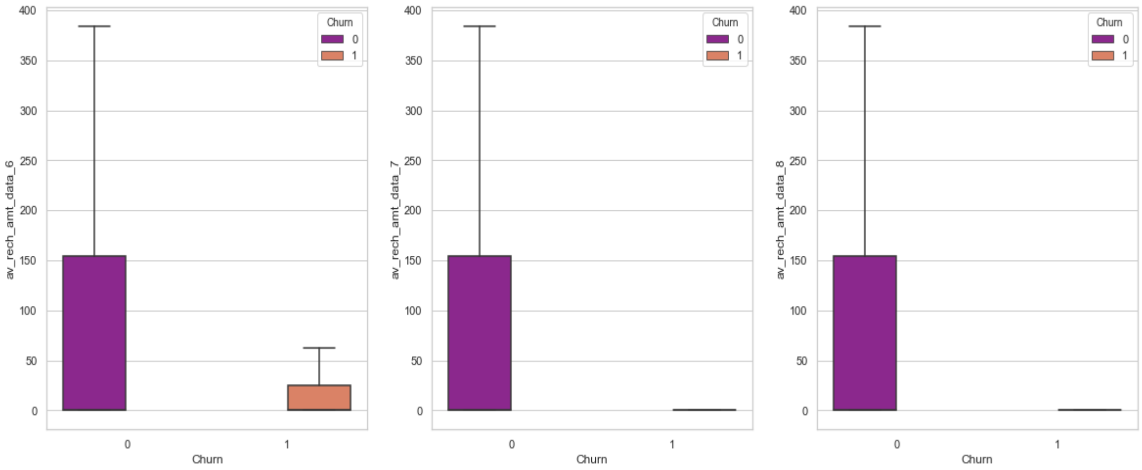


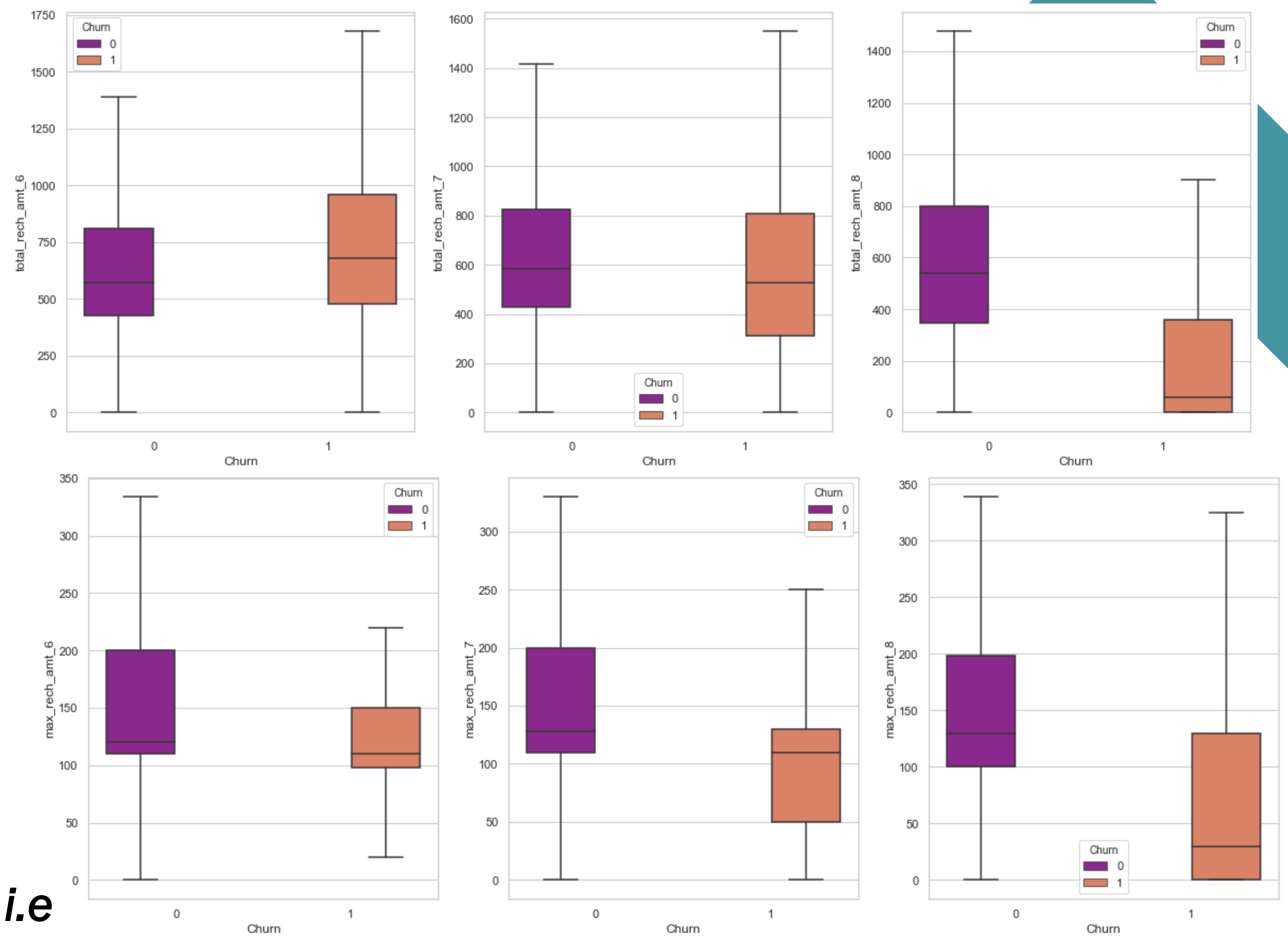




01. Data cleaning and preparation

Exploratory Data analysis data visualizations

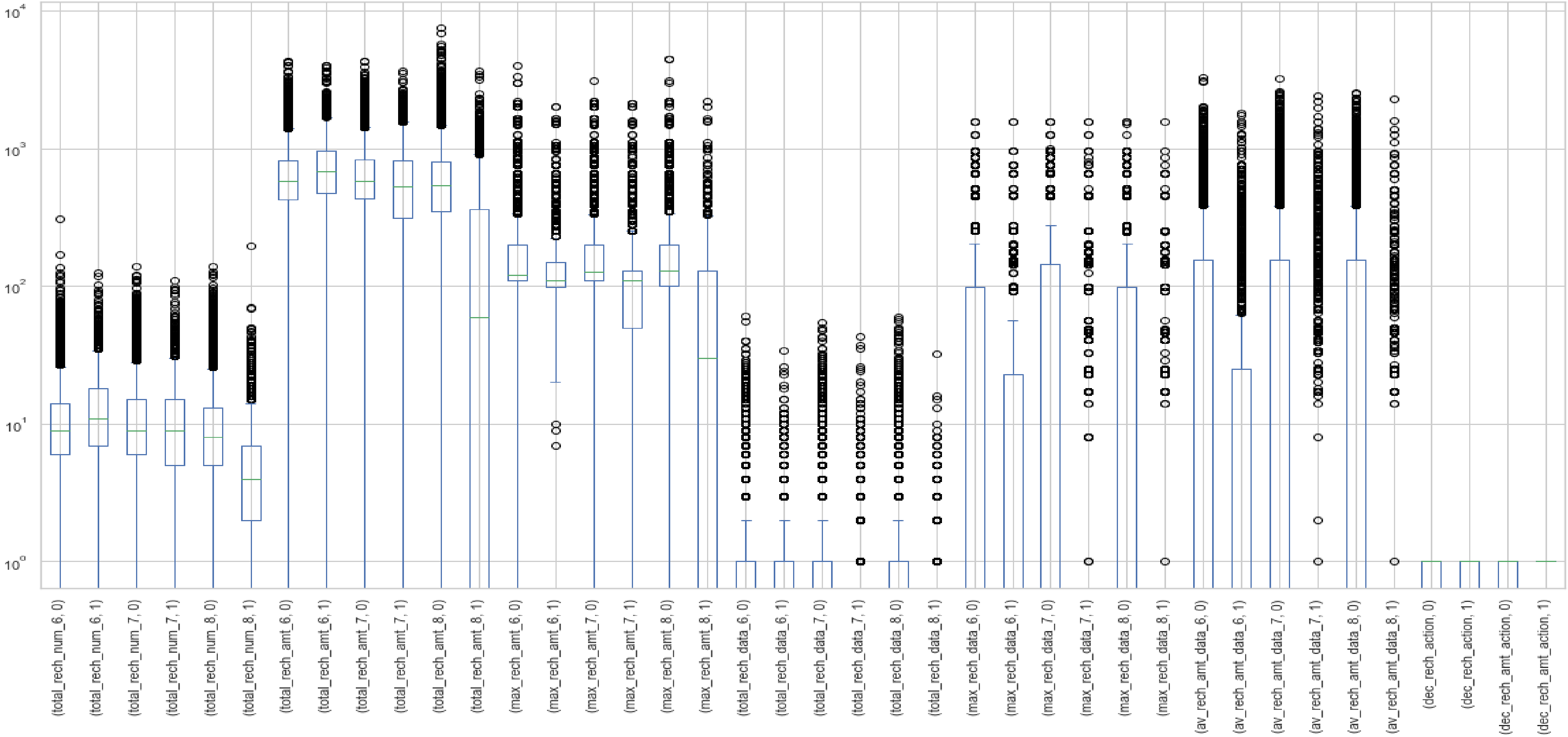




From the above plots we can see clearly that the recharge amounts (Total & Maximum) started to fall in the month 8 i.e near to the churn phase

01. Data cleaning and preparation

Exploratory Data analysis
data visualizations



02. Test-train split and Scaling

Test-train split

Total data split in 70 to 30 ratio

i.e., train size: 0.70 and test size: 0.30

Using SMOTE

Synthetic Minority Oversampling Technique for class imbalance

Train data shape: 38004 rows and 137 columns

Scaling the train data

Total scalable columns: 137

Scalable cols after removing : 133

```
[ ] X_train.head()
```

	arpu_6	arpu_7	arpu_8	onnet_mou_6	onnet_mou_7	onnet_mou_8	offnet_mou_6	offnet_mou_7	offnet_mou_8	roam_ic_mou_6	roam_ic_mou_7	roam_ic_mou_8	roam_og_mo
0	2.42	1.62	3.55	0.23	0.11	1.02	0.81	0.89	1.40	0.52	0.29	1.41	(
1	-0.60	0.02	0.82	-0.66	-0.62	-0.37	-0.48	-0.35	0.05	-0.22	-0.24	-0.24	-(
2	-1.69	0.21	0.91	-0.68	1.61	2.13	-0.85	1.14	1.97	-0.22	-0.24	-0.24	-(
3	0.84	1.23	1.62	0.48	0.15	0.49	1.44	1.58	1.98	-0.22	-0.24	-0.24	-(
4	-0.37	-0.29	0.13	-0.45	-0.32	-0.04	-0.73	-0.71	-0.59	1.41	7.44	5.89	-(

03. Model Building

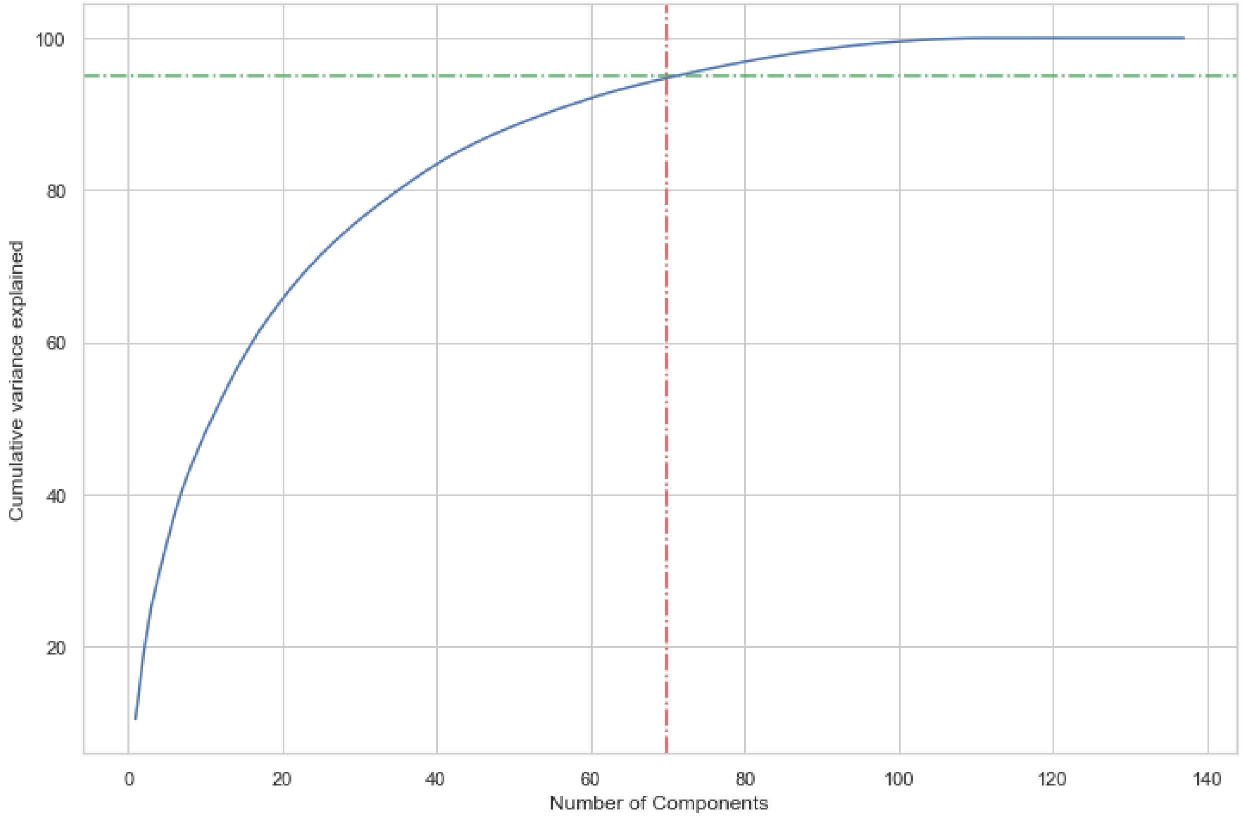
Model building with PCA

Observation: 70 components are seen enough to describe 95% of the variance in the dataset. We'll choose 70 components for our modeling

Using incremental PCA for better efficiency

max positive corr: 0.016 , min negative corr: -0.017

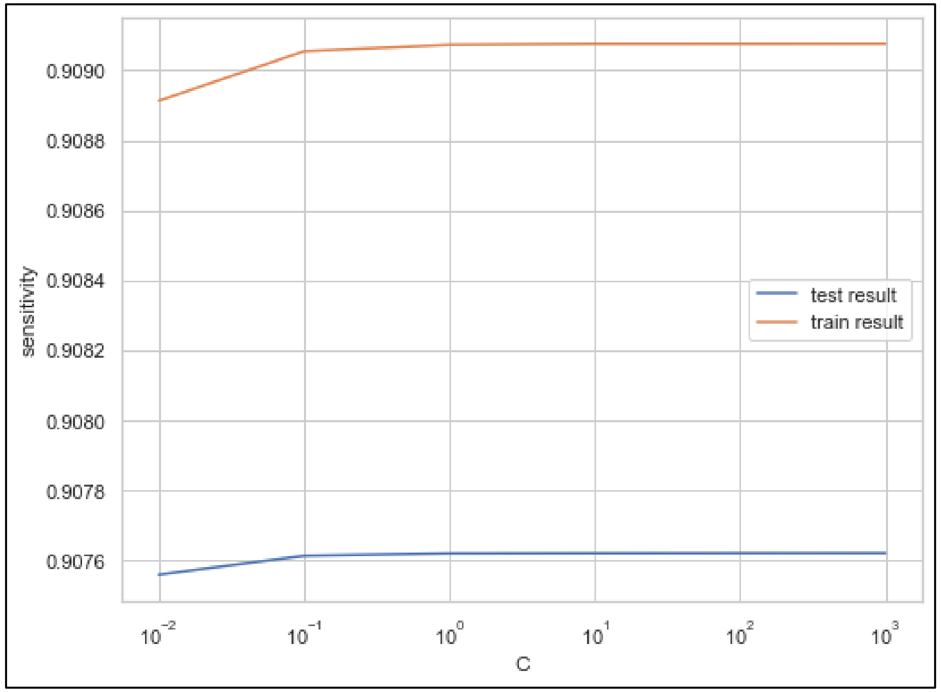
It is observed from calculations that the correlation among the attributes is almost 0, hence we proceeded with these principal components.

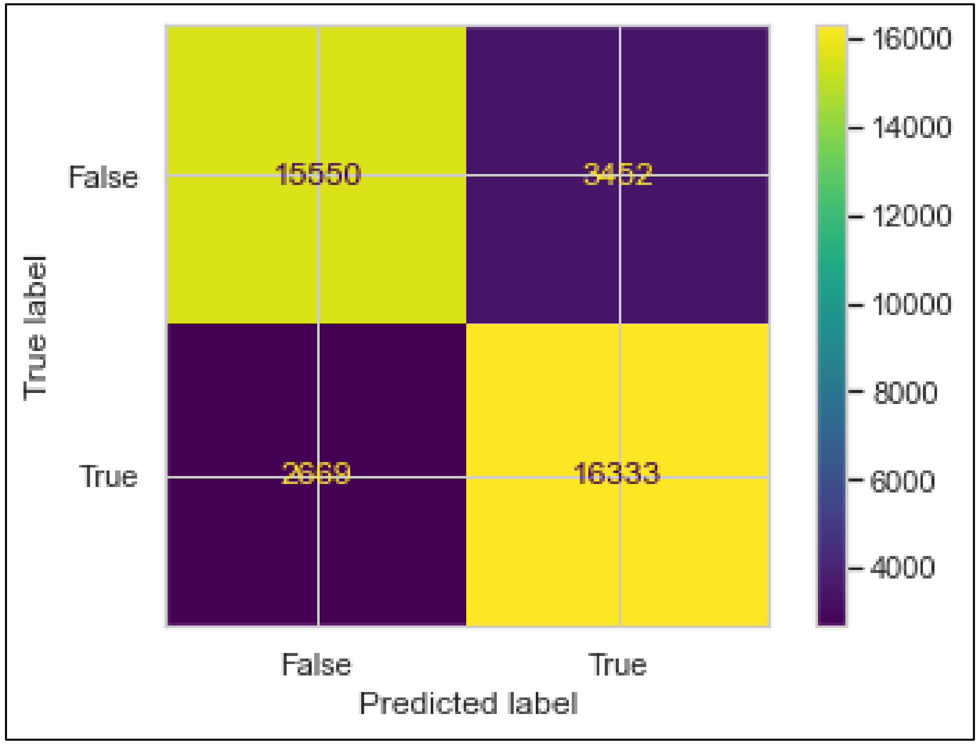


03. Model Building

Logistic regression with PCA

```
confusoin matrix
 [[6696 1447]
 [ 139  628]]
sensitivity      0.82
specificity      0.82
area under the curve     0.82
```





04. Model Evaluation

Decision tree with PCA

Model summary (Decesion Trees with PCA)

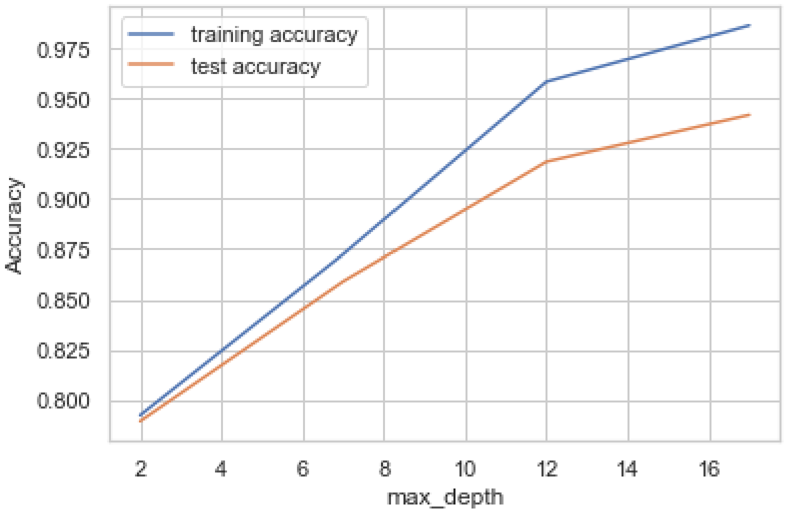
- `Train set`
  - Accuracy = 0.87
  - Sensitivity = 0.89
  - Specificity = 0.86
  - roc_auc_score= 0.87
- `Test set`
  - Accuracy = 0.83
  - Sensitivity = 0.89
  - Specificity = 0.86
  - roc_auc_score= 0.77

Sesitivity and Specificity are same while evaluating the model on the test set and Train Set and the accuracy also remained close.

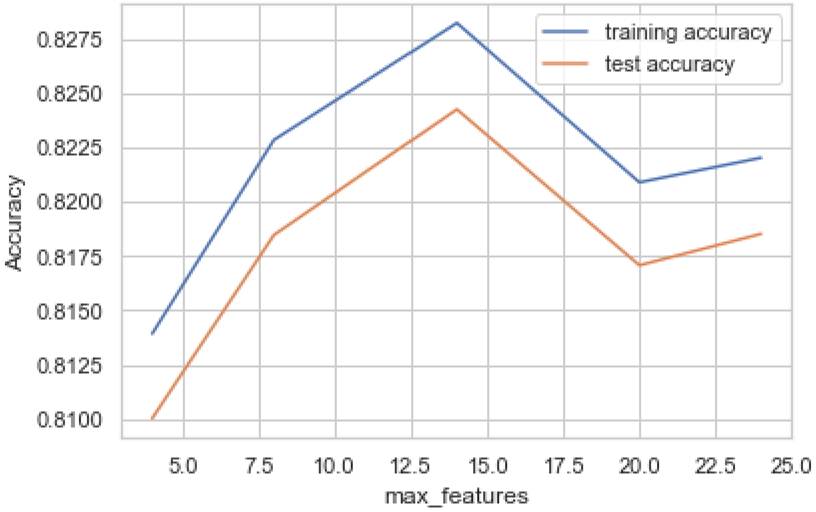
05. Prediction on test set

Random forest with PCA

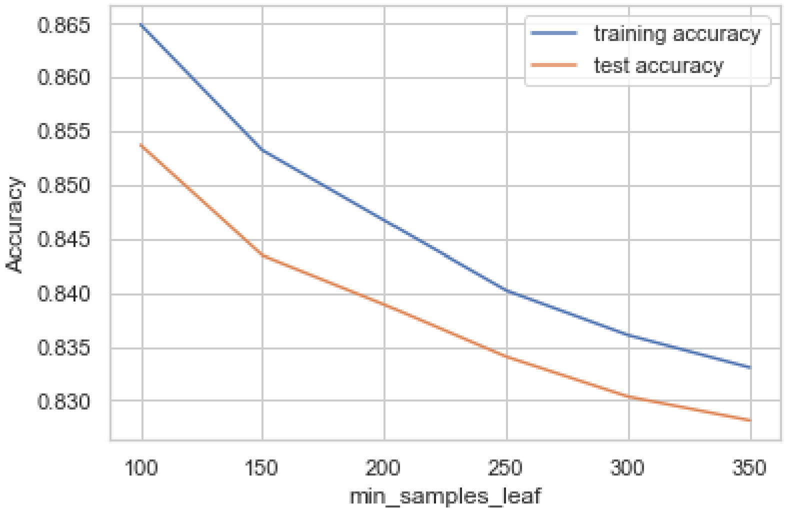
plotting accuracies with max_depth



plotting accuracies with max_features



plotting accuracies with sample leaf



Results Summary

Business insights generated

- Average revenue per customer in the 7th month was the deciding factor for the churn rate. A sharp decline indicated the customer might churn.
- Total minutes of usage for outgoing is also an important churn predictor.
- Incoming and outgoing Roaming Minutes of usage is also seen impacting churn
- The outgoing Local Minutes of usage are very crucial features on the customer churn.
- In the 8th month which is the Action Phase, there was a considerable drop in recharge

Results Summary

Strategies to reduce churn rate

- Special offers for high valued clients on recharge amounts
- Provision of special packages with special roaming rates/STD and ISD packages in the action phase
- Data package offers may not provide enough incentive to clients
- Customer satisfaction survey from time to time to keep tabs on services efficiency

Thank you