Telecom Churn Case Study

Presentation by

Siddhant Dunung

Parth Dalvi

Abisha Jotheesh Bell M P

INTRODUCTION - Business problem overview

In the 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. Given the fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, customer retention has now become even more important than customer acquisition.

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, you 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.

Objective

Predict churn risk and identify main indicators, especially among high-value customers – top 20% revenue generating customers.

Understanding the data

The dataset contains customer-level information for a span of four consecutive months - June, July, August and September. The months are encoded as 6, 7, 8 and 9, respectively.
The business objective is to predict the churn in the last (i.e. the ninth) month using the data (features) from the first three months. To do this task well, understanding the typical customer behavior during churn will be helpful.

UNDERSTANDING DATA

Load the DatasetRemoved the MissingValues from DatasetFiltered the High ValueCustomer based on therecharge amount duringAction Phase(Month of Juneand July)Tagging of Churn done by usingthe Minute of use variable andInternet use. Outliers Removed from theDataset for Numeric Variable

MODEL BUILDING

Feature Engineering donefor deriving New Variablesfor Analysis. Data is divided into Trainand Test Data Set· Data Set is imbalanced doSMOTE technique is usedfor removing imbalance. Logistic Regression Modelbuilding done usingRecursive Feature Selection

MODEL EVALUATION

Final Model after low p-value and low VIF is testedon the Train Data Set.. Accuracy and Recall Valuesobtained were good for theTest Dataset.. Model was tested on the TestDataset. Obtained Accuracyand Recall are good. SoModel is performing well onthe Test Dataset as well.

Data Cleaning

Handling Missing Values in Columns. Deleting Date Columns as they are not relevant for Analysis. Deleting columns with 1 unique values as they are not relevant for our analysis

```python
# Columns with more than 30 Percent Missing Values:
columns_to_delete=null_values_df[null_values_df.null_percentage>30].index
columns_to_delete
```

```
Index(['date_of_last_rech_data_6', 'date_of_last_rech_data_7',
       'date_of_last_rech_data_8', 'date_of_last_rech_data_9',
       'total_rech_data_6', 'total_rech_data_7', 'total_rech_data_8',
       'total_rech_data_9', 'max_rech_data_6', 'max_rech_data_7',
       'max_rech_data_8', 'max_rech_data_9', 'count_rech_2g_6',
       'count_rech_2g_7', 'count_rech_2g_8', 'count_rech_2g_9',
       'count_rech_3g_6', 'count_rech_3g_7', 'count_rech_3g_8',
       'count_rech_3g_9', 'av_rech_amt_data_6', 'av_rech_amt_data_7',
       'av_rech_amt_data_8', 'av_rech_amt_data_9', 'arpu_3g_6', 'arpu_3g_7',
       'arpu_3g_8', 'arpu_3g_9', 'arpu_2g_6', 'arpu_2g_7', 'arpu_2g_8',
       'arpu_2g_9', 'night_pck_user_6', 'night_pck_user_7', 'night_pck_user_8',
       'night_pck_user_9', 'fb_user_6', 'fb_user_7', 'fb_user_8', 'fb_user_9'],
      dtype='object')
```

Filter High Value Customers

Data cleaning for missing values in rows.
Data cleaning for missing values in rows

```python
# 70th Percentiles of avg amount recharged in good phase
telecom_df['avg_rech_amt_6_7']=(telecom_df['total_rech_amt_6']+telecom_df['total_rech_amt_7'])/2
```

```python
X=telecom_df['avg_rech_amt_6_7'].quantile(0.70)
X
```

```
368.5
```

```python
telecom_df=telecom_df[telecom_df['avg_rech_amt_6_7']>=X]
telecom_df.shape
```

```
(30011, 167)
```

Extracting the Churn Variable

```
telecom_df['Churn']=(telecom_df['total_ic_mou_9']+telecom_df['total_og_mou_9']+telecom_df['vol_2g_mb_9']+telecom_df['vol_3g_mb_9']).map(lambda X:1 if X==(
telecom_df.head()
```

	mobile_number	arpu_6	arpu_7	arpu_8	arpu_9	onnet_mou_6	onnet_mou_7	onnet_mou_8	onnet_mou_9	offnet_mou_6	offnet_mou_7	offnet_mou_8	offnet_mo
8	7001524846	378.721	492.223	137.362	166.787	413.69	351.03	35.08	33.46	94.66	80.63	136.48	10
13	7002191713	492.846	205.671	593.260	322.732	501.76	108.39	534.24	244.81	413.31	119.28	482.46	21
16	7000875565	430.975	299.869	187.894	206.490	50.51	74.01	70.61	31.34	296.29	229.74	162.76	22
17	7000187447	690.008	18.980	25.499	257.583	1185.91	9.28	7.79	558.51	61.64	0.00	5.54	8
21	7002124215	514.453	597.753	637.760	578.596	102.41	132.11	85.14	161.63	757.93	896.68	983.39	86

Removing Data for Churn Phase and dropping values for churn phase

```
# Dropping all values for Churn phase
telecom_df=telecom_df.drop(column_9,axis=1)
```

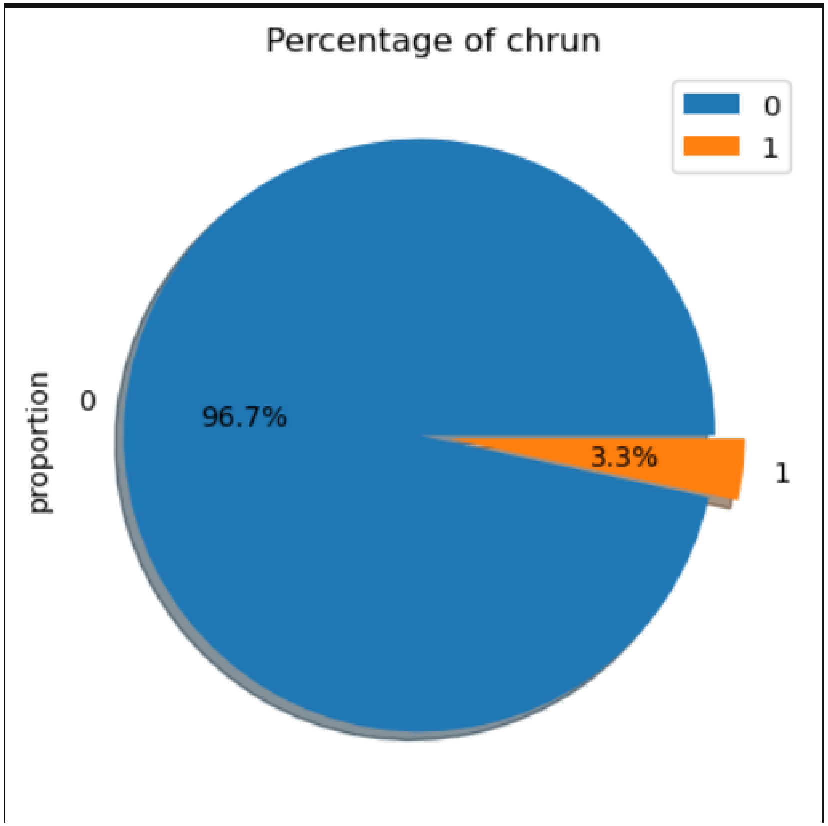
```
# Dropping column for churn phase
telecom_df=telecom_df.drop('sep_vbc_3g',axis=1)
```

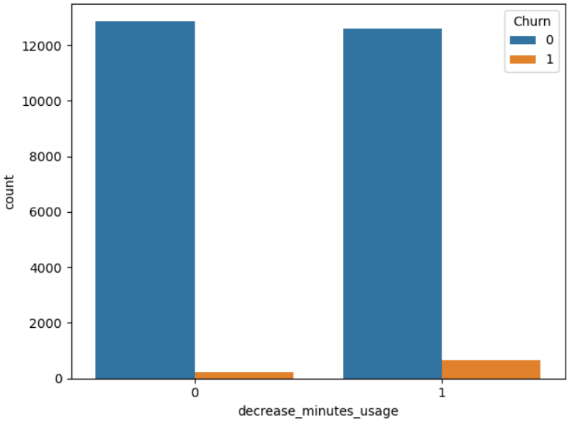
```
telecom_df.shape
```

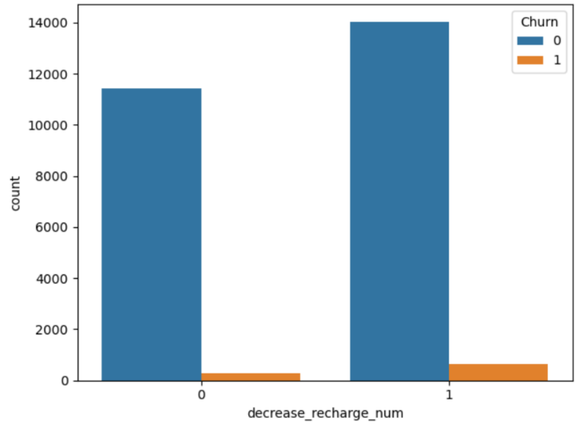
```
(27991, 127)
```

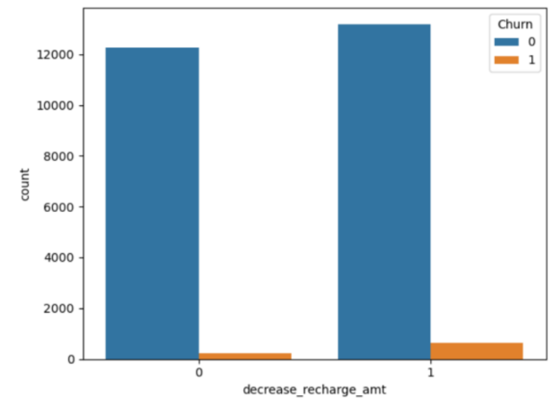
Exploratory Data Analysis

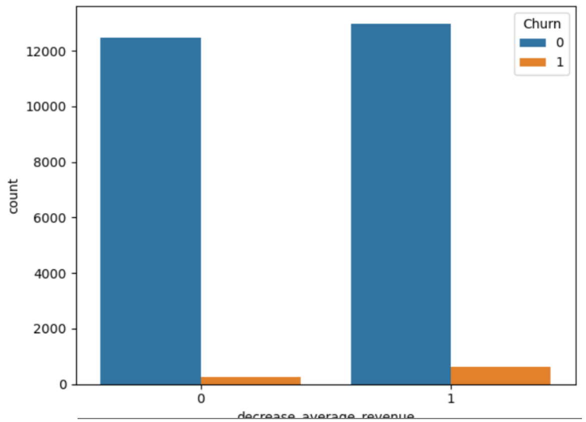
While checking the churn percentage, We can infer from calculations and pie chart that is a case of class imbalance

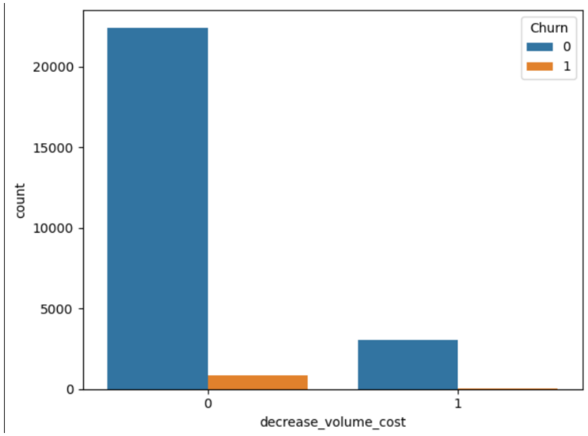


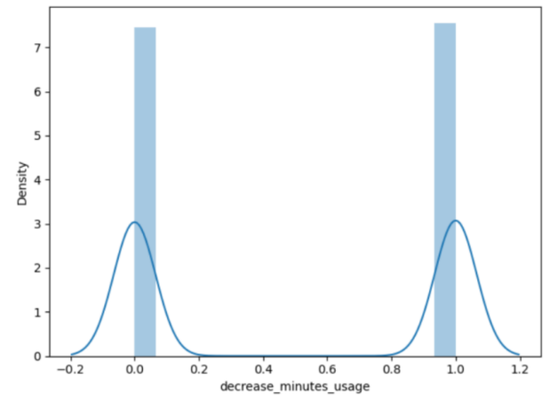

Churn
0
1
count
decrease_minutes_usage

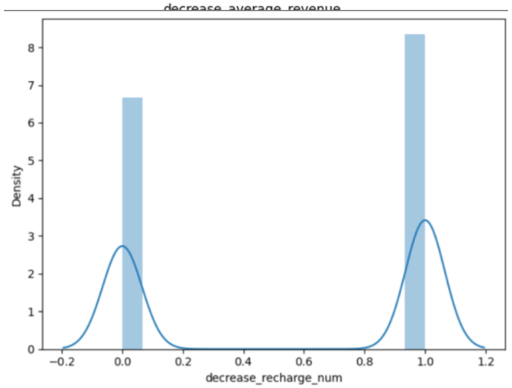

Churn
0
1
count
decrease_recharge_num

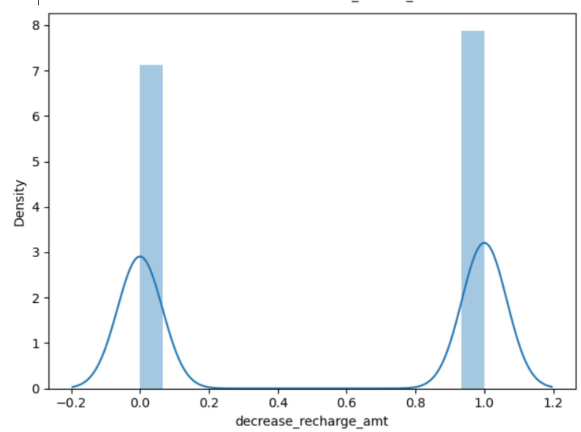

Churn
0
1
count
decrease_recharge_amt

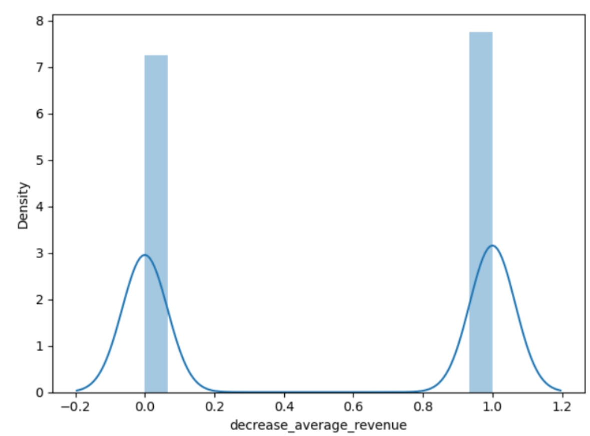

Churn
0
1
count

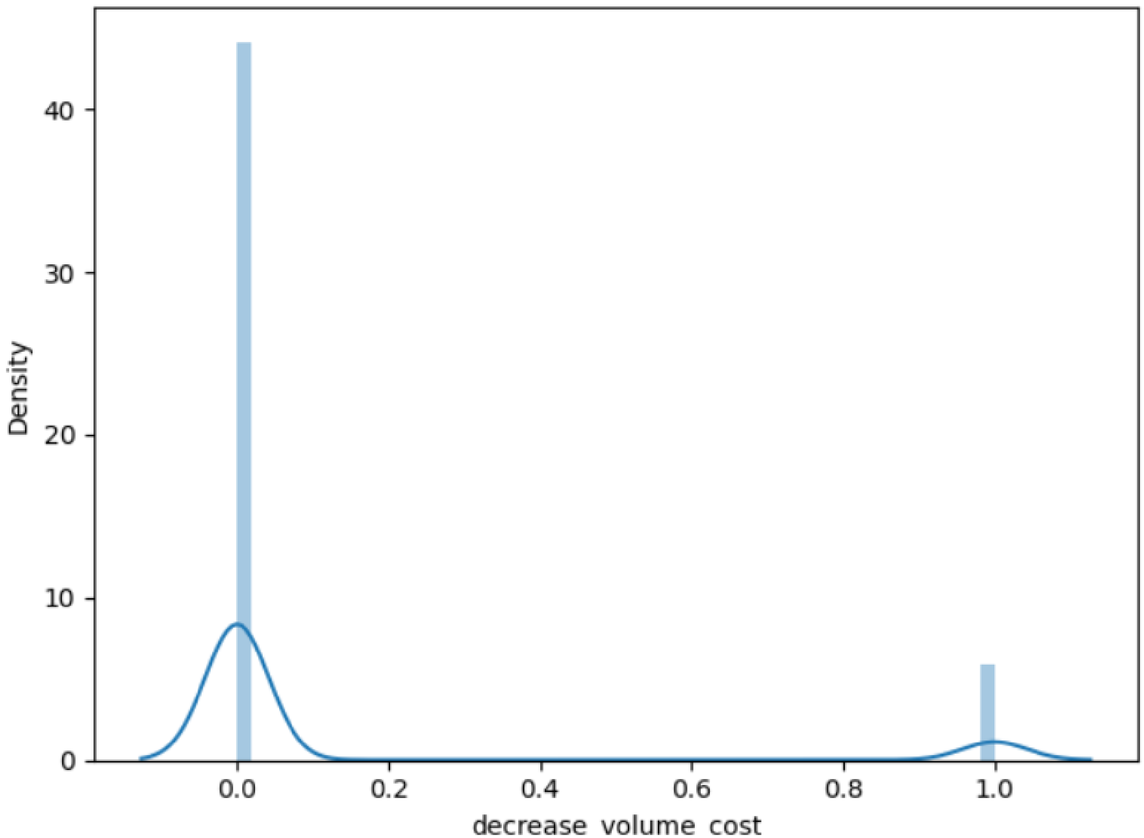

Churn
0
1
count
decrease_volume_cost


Density
decrease_minutes_usage

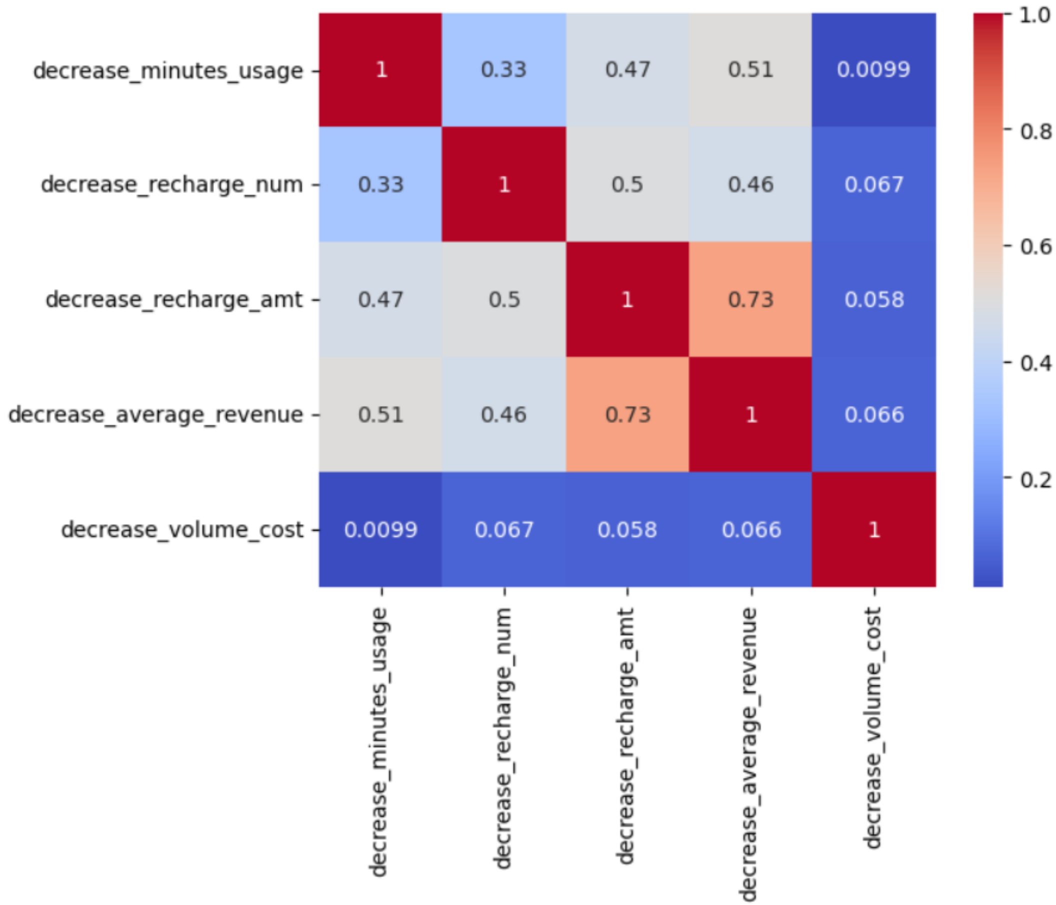

Density
decrease_recharge_num


Density
decrease_recharge_amt


Density
decrease_average_revenue



- *Churn is high for customers which have reduced their minutes of usage in the action phase.*
- *Churn is high for customer who have reduced their recharge number in action phase.*
- *Customers who have decreased the recharge number in action phase are more prone to churn.*
- *customers whose volume based cost is increased are more vulnerable to churn.*



There is a high correlation in decrease in recharge revenue and recharge amount. However, is not that high that we have to drop it.

Logistic Regression

Model-1

```
# Creating a variable for rf columns
X_train_columns = X_train[rfe_columns]
```

```
X_train_sm_1 =sm.add_constant(X_train_columns)
log_sm = sm.GLM(y_train,X_train_sm_1,family=sm.families.Binomial())
model_1 = log_sm.fit()
model_1.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	Churn	**No. Observations:**	35654
Model:	GLM	**Df Residuals:**	35638
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	nan
Date:	Tue, 04 Feb 2025	**Deviance:**	25576.
Time:	16:06:09	**Pearson chi2:**	4.01e+07
No. Iterations:	40	**Pseudo R-squ. (CS):**	nan
Covariance Type:	nonrobust		

Model-2

```
X_train_columns = X_train_columns.drop('total_og_mou_8',axis=1)
```

```
X_train_sm=sm.add_constant(X_train_columns)
log_sm2 =sm.GLM(y_train,X_train_sm,family=sm.families.Binomial())
model_2=log_sm2.fit()
model_2.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	Churn	**No. Observations:**	35654
Model:	GLM	**Df Residuals:**	35639
Model Family:	Binomial	**Df Model:**	14
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	nan
Date:	Tue, 04 Feb 2025	**Deviance:**	25629.
Time:	16:06:14	**Pearson chi2:**	3.30e+07
No. Iterations:	41	**Pseudo R-squ. (CS):**	nan
Covariance Type:	nonrobust		

Model-3

```
X_train_columns = X_train_columns.drop('offnet_mou_8',axis=1)
```

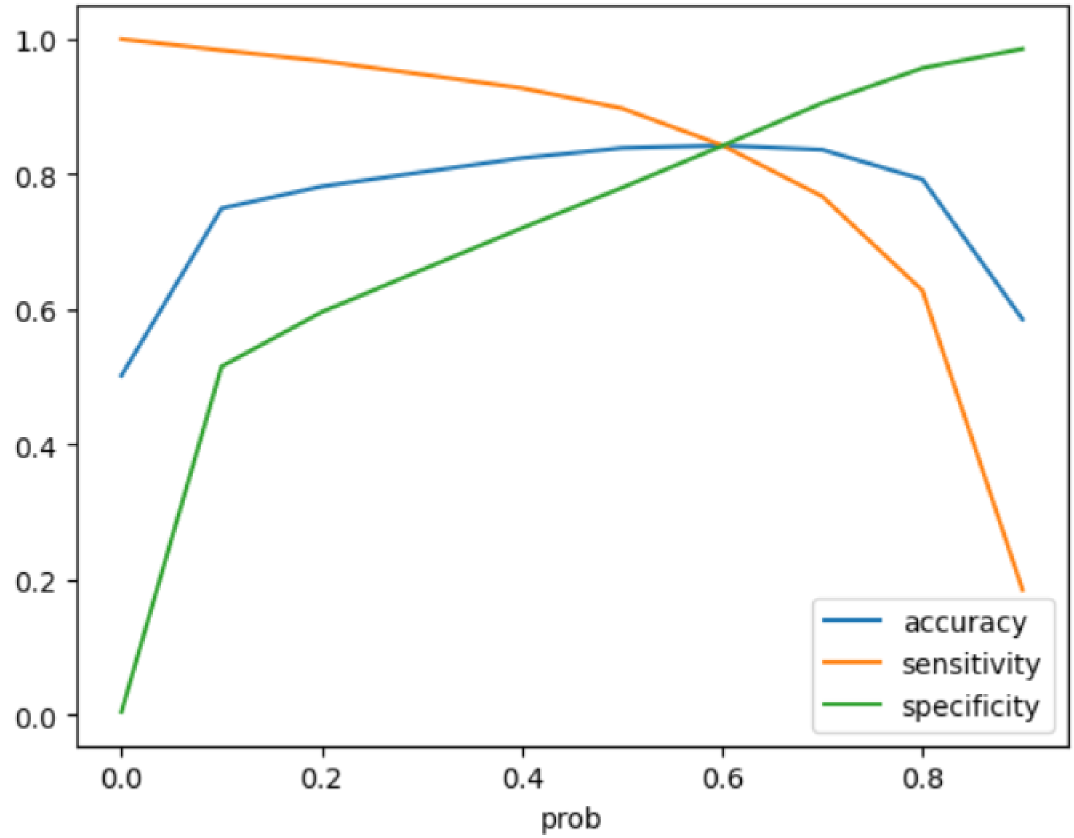
```
X_train_sm_3=sm.add_constant(X_train_columns)
log_sm_3 =sm.GLM(y_train,X_train_sm_3,family=sm.families.Binomial())
model_3=log_sm_3.fit()
model_3.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	Churn	**No. Observations:**	35654
Model:	GLM	**Df Residuals:**	35640
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	nan
Date:	Tue, 04 Feb 2025	**Deviance:**	26740.
Time:	16:06:18	**Pearson chi2:**	1.88e+07
No. Iterations:	40	**Pseudo R-squ. (CS):**	nan
Covariance Type:	nonrobust		

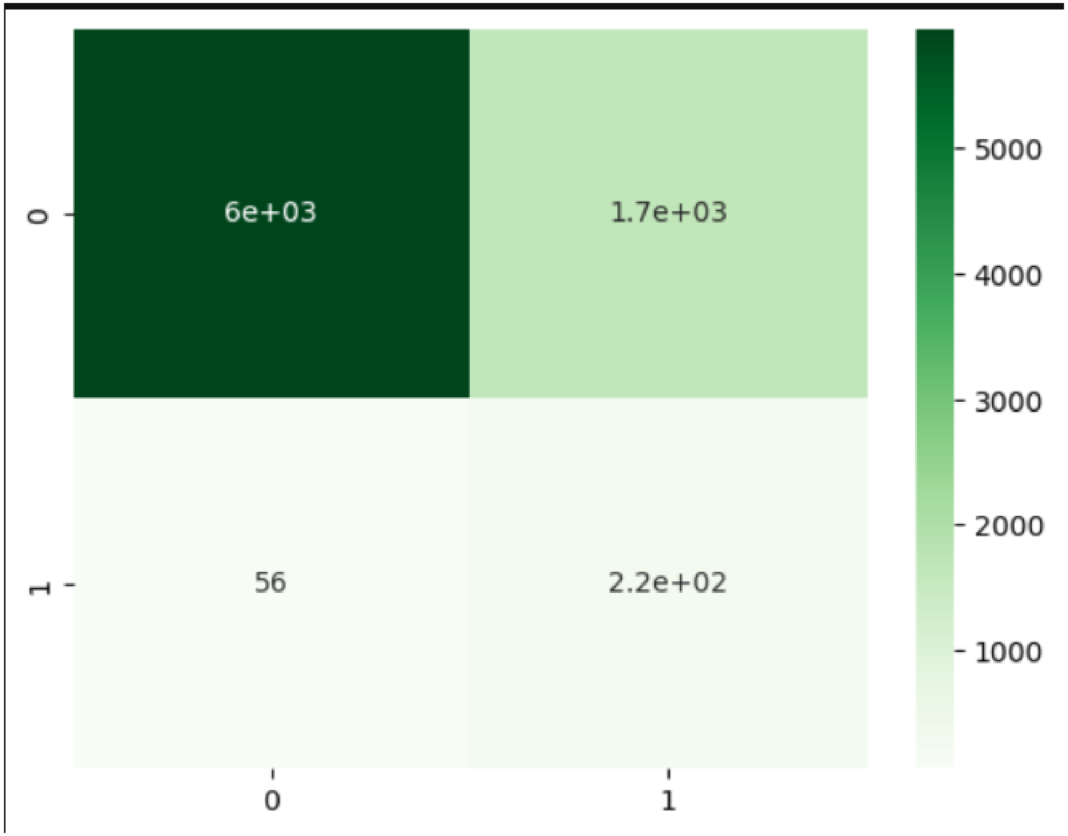
The data is imbalanced and we treated it with Synthetic Minority Oversampling Technique

Evaluation



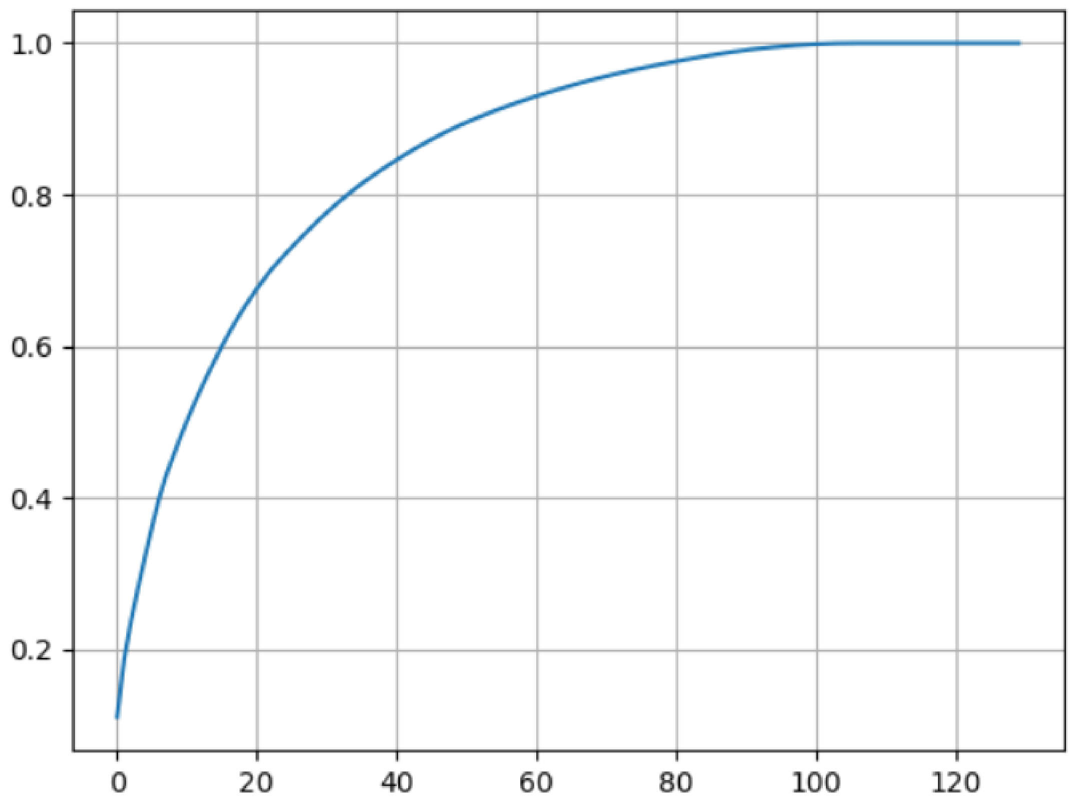
Cut off Point is 0.6 at which Sensitivity, Specificity, Accuracy is stable

Evaluation of model on test set



From above mentioned Metrics we can conclude that overall performance on Test set is good

Performing Model with Principal Component Analysis



Evaluation of model on Train Data

```
y_train_pred_dt = decision_model.predict(X_train_pca)
```

```
Confusion_matrix = confusion_matrix(y_train, y_train_pred_dt)
print(Confusion_matrix)
```

```
[[15590  2237]
 [ 1255 16572]]
```

```
TP = Confusion_matrix[1,1] # true positive
TN = Confusion_matrix[0,0] # true negatives
FP = Confusion_matrix[0,1] # false positives
FN = Confusion_matrix[1,0] # false negatives
```

```
print('Accuracy', accuracy_score(y_train, y_train_pred_dt))
print('Precision', precision_score(y_train, y_train_pred_dt))
print('Recall', recall_score(y_train, y_train_pred_dt))
print("Sensitivity:-",TP / float(TP+FN))
print("Specificity:-", TN / float(TN+FP))
```

```
Accuracy 0.9020586750434734
Precision 0.8810675740337073
Recall 0.9296011667695069
Sensitivity:- 0.9296011667695069
Specificity:- 0.8745161833174399
```

Evaluation of model on test data

```
y_test_pred_dt = decision_model.predict(X_test_pca)
```

```
Confusion_matrix = confusion_matrix(y_test,y_test_pred_dt)
Confusion_matrix
```

```
array([[4634, 2986],
       [ 103,  176]], dtype=int64)
```

```
print('Accuracy', accuracy_score(y_test,y_test_pred_dt))
print('Precision', precision_score(y_test,y_test_pred_dt))
print('Recall', recall_score(y_test,y_test_pred_dt))
print("Sensitivity:-",TP / float(TP+FN))
print("Specificity:-", TN / float(TN+FP))
```

```
Accuracy 0.6089378402329408
Precision 0.05566097406704617 6
Recall 0.6308243727598566
Sensitivity:- 0.9296011667695069
Specificity:- 0.8745161833174399
```

Conclusion/Suggestions

Company must provide some reward to long term and high value customers in order to retain them.
Company must offer rates and offers in such a way that it matches with the competitors.
Company must encourage user to use more data packages such as free data at night, free vouchers on top up, combo offers with OTT subscriptions like Netflix etc.
Company must take feedback from customers in order to understand their needs and avoid churn.
Company should also focus on network connectivity especially in those areas where majority of customers reside.