TELECOM CHURN CASE STUDY

INTRODUCTION

- **Churn prediction** is one of the most popular Big Data use cases in Business. It consists of detecting customers who are likely to cancel a subscription to a service.

- **Churn** is a problem for telecom companies because it is more expensive to acquire a more customers than to keep your existing one from leaving.

PROBLEM STATEMENT

- 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.

UNDERSTANDING & DEFINING CHURN

- There are two main models of payment in the telecom industry - **postpaid** (customers pay a monthly/annual bill after using the services) and **prepaid** (customers pay/recharge with a certain amount in advance and then use the services).
- In the postpaid model, when customers want to switch to another operator, they usually inform the existing operator to terminate the services, and you directly know that this is an instance of churn.
- However, in the prepaid model, customers who want to switch to another network can simply stop using the services without any notice, and it is hard to know whether someone has actually churned or is simply not using the services temporarily (e.g. someone may be on a trip abroad for a month or two and then intend to resume using the services again).
- Thus, churn prediction is usually more critical (and non-trivial) for prepaid customers, and the term ‘churn’ should be defined carefully. Also, prepaid is the most common model in India and Southeast Asia, while postpaid is more common in Europe in North America.
- This project is based on the Indian and Southeast Asian market.

DEFINITIONS OF CHURN

- **Revenue-based churn**: Customers who have not utilised any revenue-generating facilities such as mobile internet, outgoing calls, SMS etc. over a given period of time. One could also use aggregate metrics such as ‘customers who have generated less than INR 4 per month in total/average/median revenue.
- The main shortcoming of this definition is that there are customers who only receive calls/SMSes from their wage-earning counterparts, i.e. they don’t generate revenue but use the services. For example, many users in rural areas only receive calls from their wage-earning siblings in urban areas.
- **Usage-based churn**: Customers who have not done any usage, either incoming or outgoing - in terms of calls, internet etc. over a period of time.
- A potential shortcoming of this definition is that when the customer has stopped using the services for a while, it may be too late to take any corrective actions to retain them. For e.g., if you define churn based on a ‘two-months zero usage’ period, predicting churn could be useless since by that time the customer would have already switched to another operator.
- In this project, you will use the **usage-based definition** to define churn.

UNDERSTANDING THE BUSINESS OBJECTIVE AND 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 behaviour during churn will be helpful.

UNDERSTANDING CUSTOMER BEHAVIOUR DURING CHURN

- Customers usually do not decide to switch to another competitor instantly, but rather over a period of time (this is especially applicable to high-value customers). In churn prediction, we assume that there are **three phases** of the customer lifecycle :
- The ‘good’ phase: In this phase, the customer is happy with the service and behaves as usual.
- The ‘action’ phase: The customer experience starts to sore in this phase, for e.g. he/she gets a compelling offer from a competitor, faces unjust charges, becomes unhappy with service quality etc. In this phase, the customer usually shows different behaviour than in the ‘good’ months. Also, it is crucial to identify high-churn-risk customers in this phase, since some corrective actions can be taken at this point (such as matching the competitor’s offer/improving the service quality etc.)
- The ‘churn’ phase: In this phase, the customer is said to have churned. You **define churn based on this phase**. Also, it is important to note that at the time of prediction (i.e. the action months), this data is not available to you for prediction. Thus, after tagging churn as 1/0 based on this phase, you discard all data corresponding to this phase.
- In this case, since you are working over a four-month window, the first two months are the ‘good’ phase, the third month is the ‘action’ phase, and the fourth month is the ‘churn’ phase.

METHODOLOGY

- Data cleaning and data manipulation.
    1. Check and handle duplicate data.
    2. Check and handle NA values and missing values.
    3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
    4. Imputation of the values, if necessary.
    5. Check and handle outliers in data.
- EDA

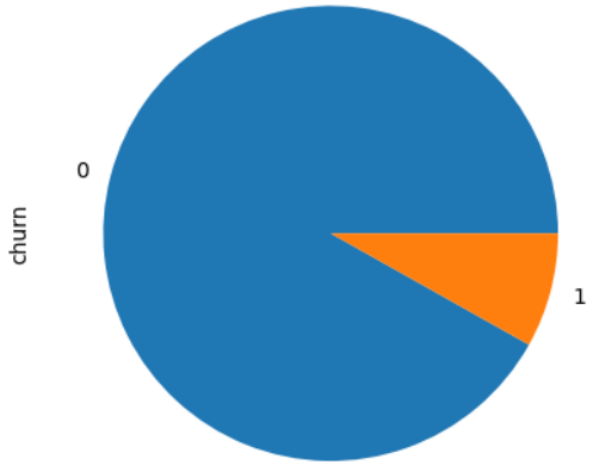
1. Univariate data analysis: value count, distribution of variable etc.
2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

EXPLORATORY DATA ANALYSIS

```
|: #Lets find churn / non churn percentage

print((df['churn'].value_counts()/len(df))*100)
((df['churn'].value_counts()/len(df))*100).plot(kind="pie")
plt.show()
```

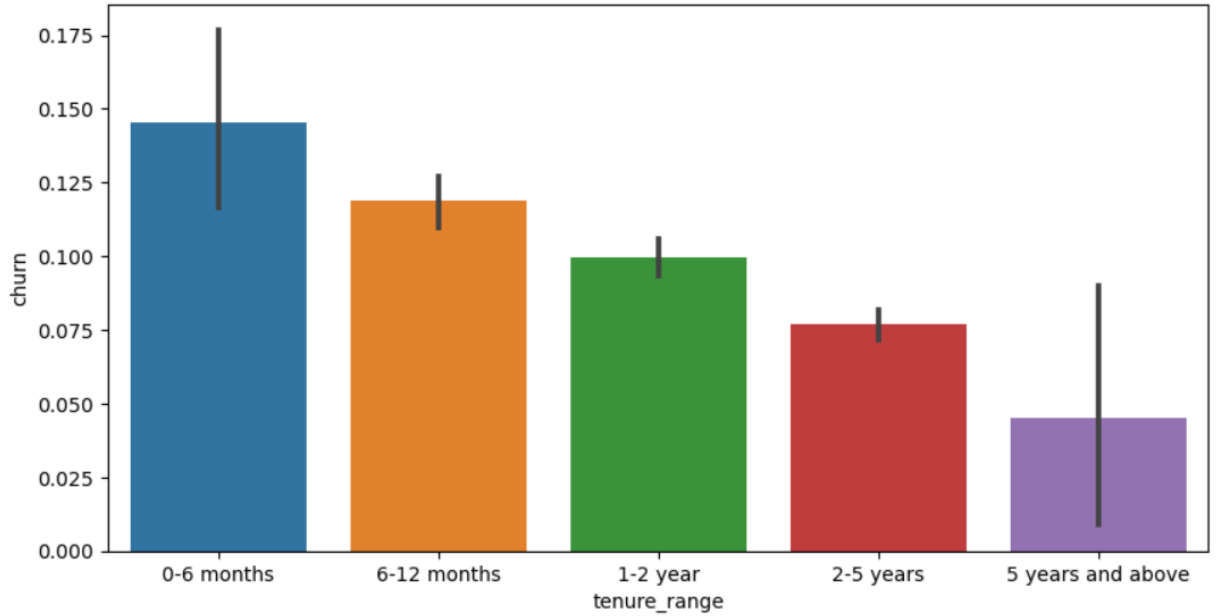
```
0    91.863605
1    8.136395
Name: churn, dtype: float64
```



As we can see that 92% of the customers do not churn, there is a possibility of class imbalance

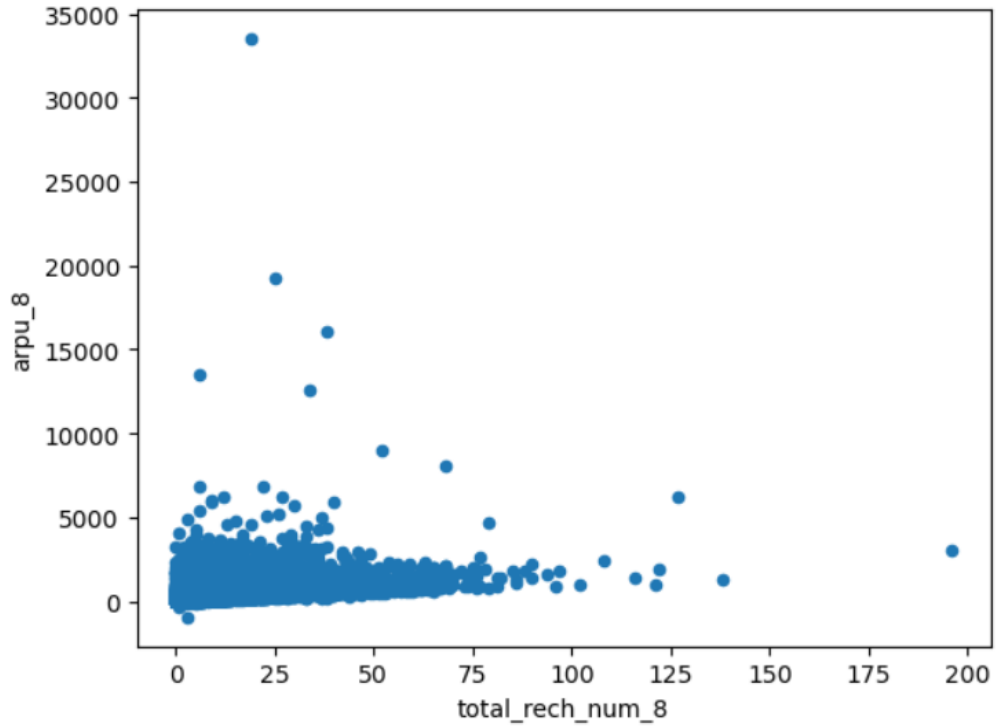
Plotting bar plot for tenure range

```
#plotting bar plot for tenure range

plt.figure(figsize=(10,5))
sns.barplot(x='tenure_range', y='churn', data=df)
plt.show()
```

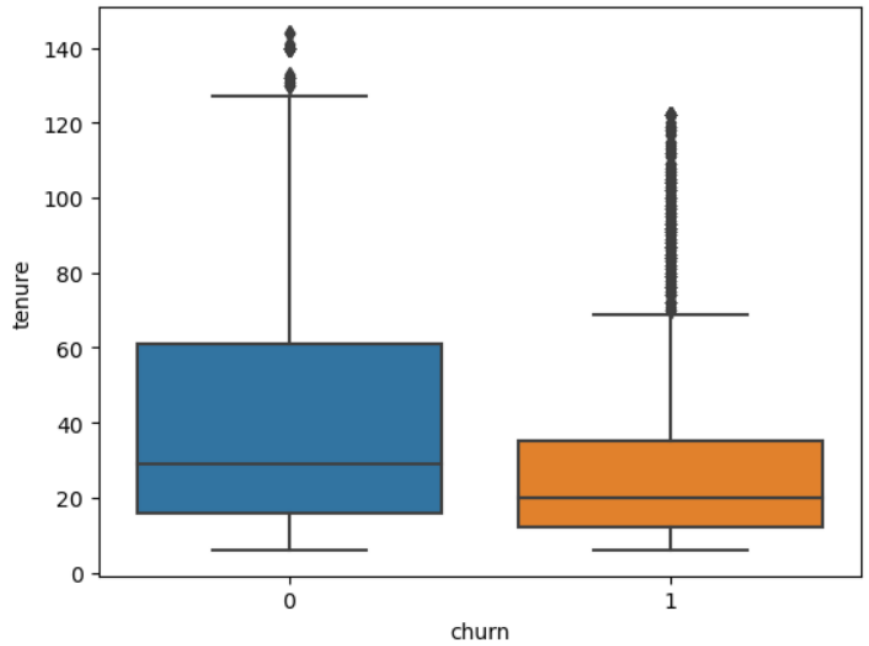


It can be seen that the maximum churn rate happens within 0-6 month, but it gradually decreases as the customer retains in the network

```
#lets plot scatter plot between total recharge and avg revenue

df[['total_rech_num_8', 'arpu_8']].plot.scatter(x='total_rech_num_8', y='arpu_8')
plt.show()
```



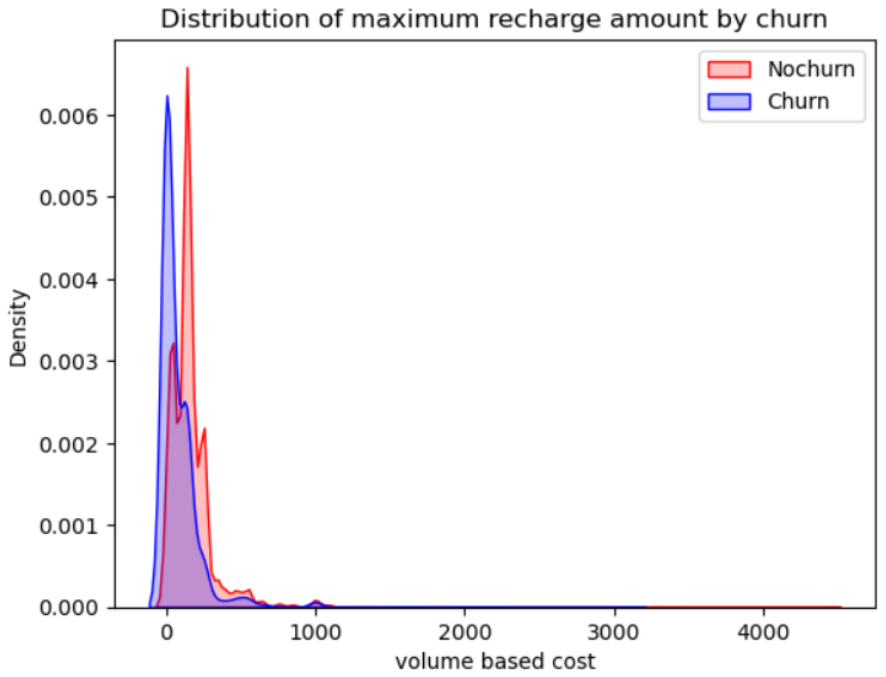
```
sns.boxplot(x = df.churn, y=df.tenure)
plt.show()
```



From the above plot , its clear tenured customers do no churn and they keep availing telecom services

Plot between churn and max rech amt

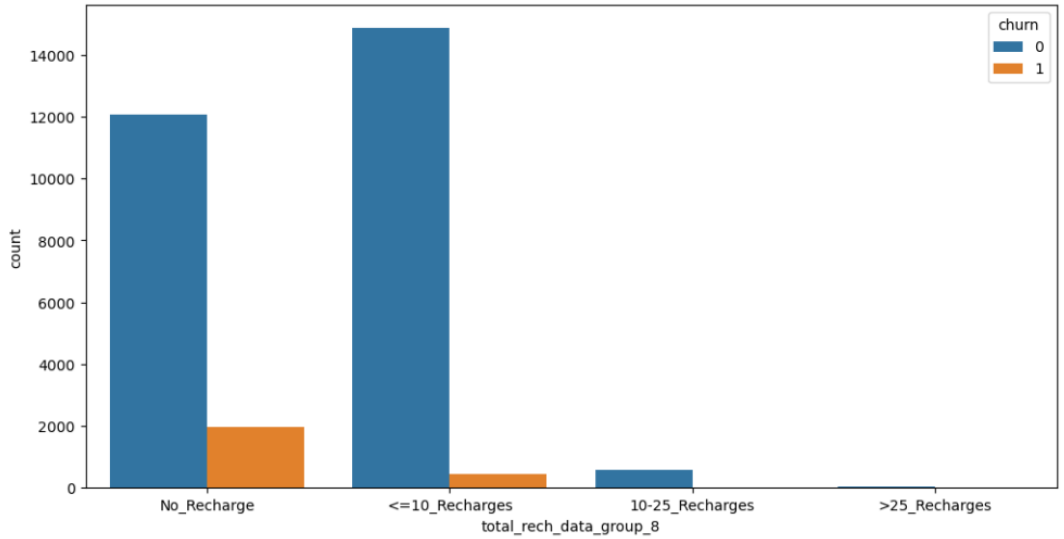
```
#plot between churn and max rech amount
ax = sns.kdeplot(df.max_rech_amt_8[(df["churn"] ==0)],color= "red", shade=True)
ax = sns.kdeplot(df.max_rech_amt_8[(df["churn"] ==1)], ax=ax, color="blue", shade=True)
ax.legend(["Nochurn", "Churn"], loc='upper right')
ax.set_ylabel('Density')
ax.set_xlabel('volume based cost')
ax.set_title('Distribution of maximum recharge amount by churn')
plt.show()
```

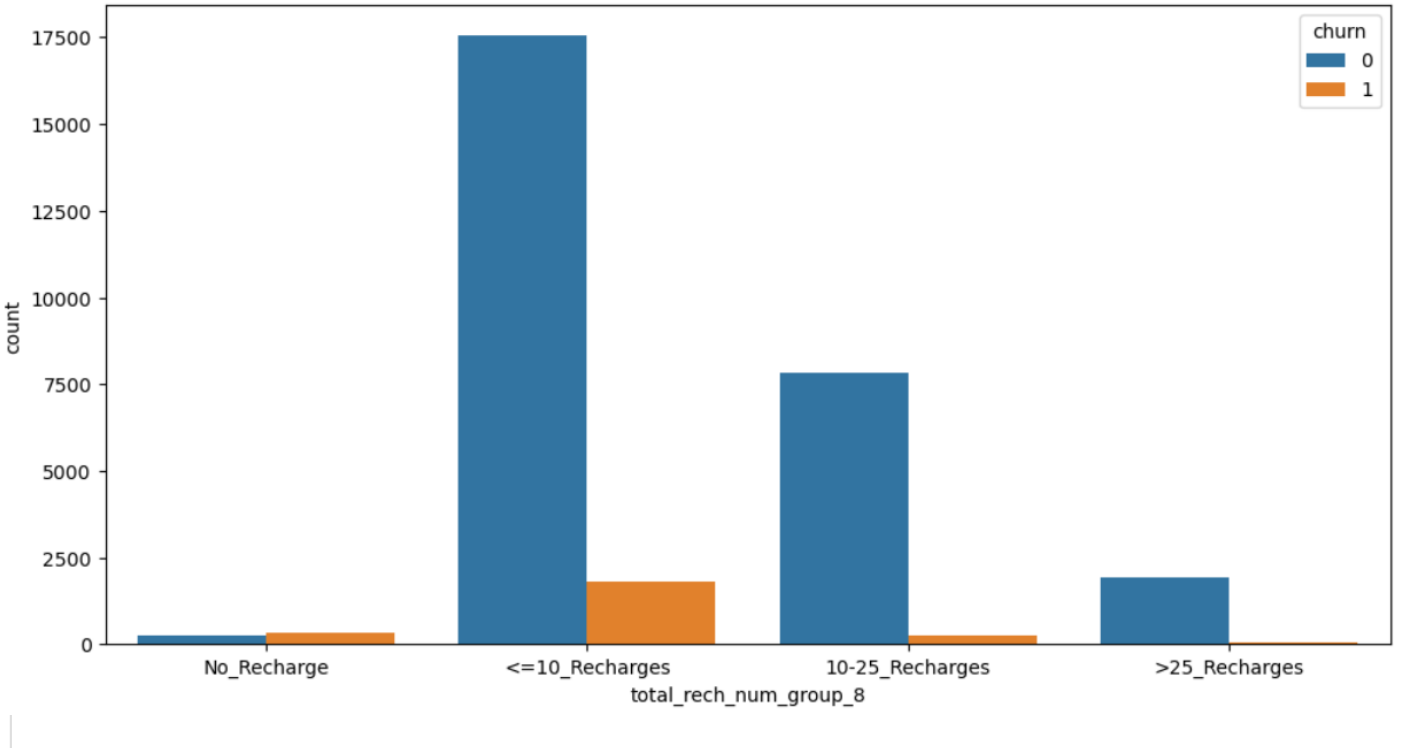


Plotting the results

```
#plotting the results
plt.figure(figsize=(12,6))
sns.countplot(data =df, x='total_rech_data_group_8', hue="churn")
print(df['total_rech_data_group_8'].value_counts())
plt.show()

plt.figure(figsize=(12,6))
sns.countplot(data =df, x='total_rech_num_group_8', hue="churn")
print(df['total_rech_data_group_8'].value_counts())
plt.show()
```

```
<=10_Recharges     15307
No_Recharge        14048
10-25_Recharges    608
>25_Recharges      38
Name: total_rech_data_group_8, dtype: int64
```





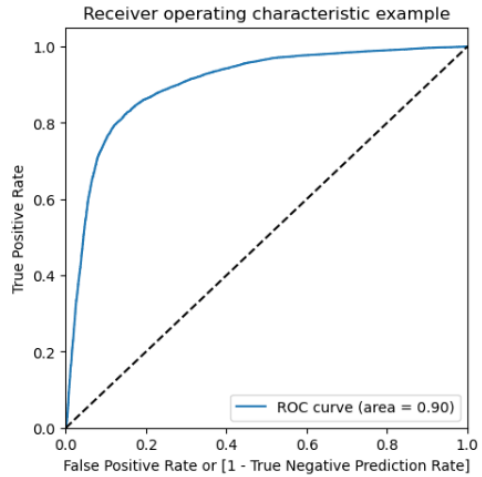
As the number of recharge rate increases, the churn rate decreases clearly.

Plotting the ROC Curve

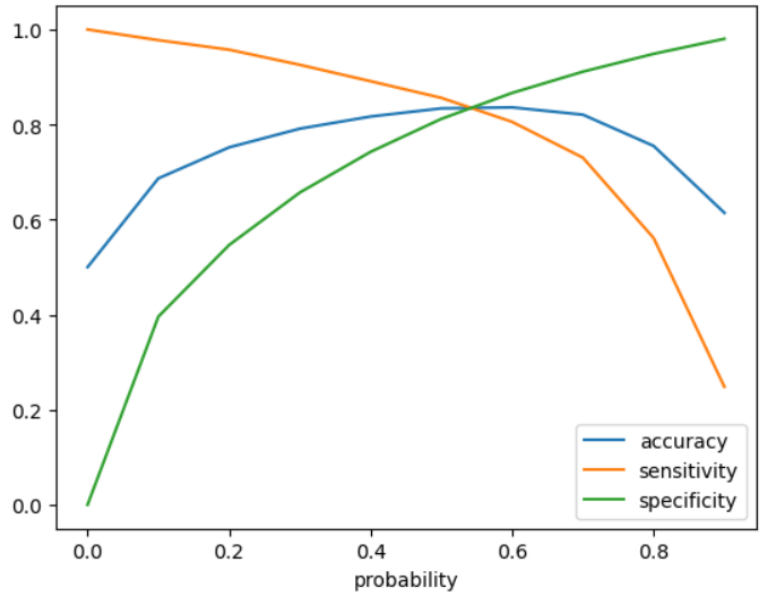
```
# Defining a function to plot the roc curve
def draw_roc( actual, probs ):
    fpr, tpr, thresholds = metrics.roc_curve( actual, probs,
                                              drop_intermediate = False )
    auc_score = metrics.roc_auc_score( actual, probs )
    plt.figure(figsize=(5, 5))
    plt.plot( fpr, tpr, label='ROC curve (area = %0.2f)' % auc_score )
    plt.plot([0, 1], [0, 1], 'k--')
    plt.xlim([0.0, 1.0])
    plt.ylim([0.0, 1.05])
    plt.xlabel('False Positive Rate or [1 - True Negative Prediction Rate]')
    plt.ylabel('True Positive Rate')
    plt.title('Receiver operating characteristic example')
    plt.legend(loc="lower right")
    plt.show()

    return None
```

```
# Defining the variables to plot the curve
fpr, tpr, thresholds = metrics.roc_curve( y_train_sm_pred_final.Converted, y_train_sm_pred_final.Converted_prob, drop_intermediat
```

```
# Plotting the curve for the obtained metrics
draw_roc(y_train_sm_pred_final.Converted, y_train_sm_pred_final.Converted_prob)
```



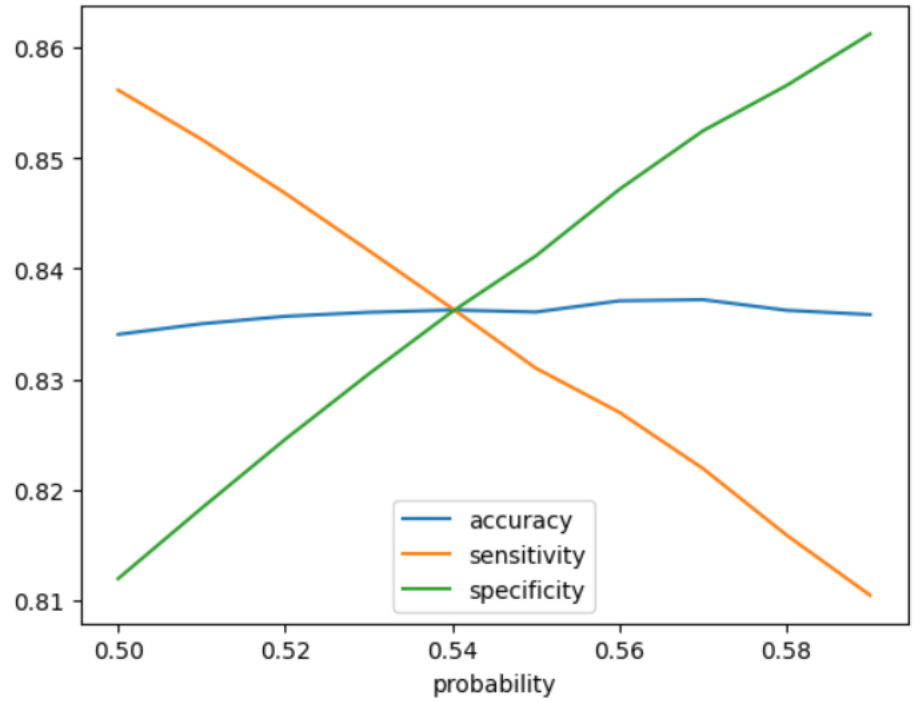
```
# plotting accuracy sensitivity and specificity for various probabilities calculated above.
cutoff_df.plot.line(x='probability', y=['accuracy','sensitivity','specificity'])
plt.show()
```



Initially we selected the optimm point of classification as 0.5.

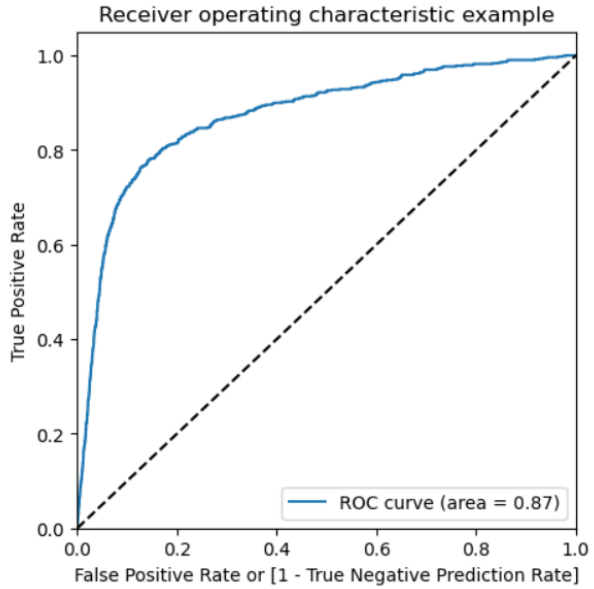
From the above graph, we can see the optimum cutoff is slightly higher than 0.5 but lies lower than 0.6. So lets tweek a little more within this range.

```
# plotting accuracy sensitivity and specificity for various probabilities calculated above.
cutoff_df.plot.line(x='probability', y=['accuracy','sensitivity','specificity'])
plt.show()
```



From the above graph we can conclude, the optimal cutoff point in the probability to define the predicted churn variabe converges at 0.54

```
# ROC curve for the test dataset

# Defining the variables to plot the curve
fpr, tpr, thresholds = metrics.roc_curve(y_pred_final.churn,y_pred_final.Conv_prob, drop_intermediate = False )
# Plotting the curve for the obtained metrics
draw_roc(y_pred_final.churn,y_pred_final.Conv_prob)
```



The AUC score for train dataset is 0.90 and the test dataset is 0.87.

This model can be considered as a good model

BUSINESS RECOMMENDATIONS

- 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.
- Customers, 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.
- Customers 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.