TELECOME CHURN CASE STUDY

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.

BUSINESS OBJECTIVE

- 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.

METHODOLOGY

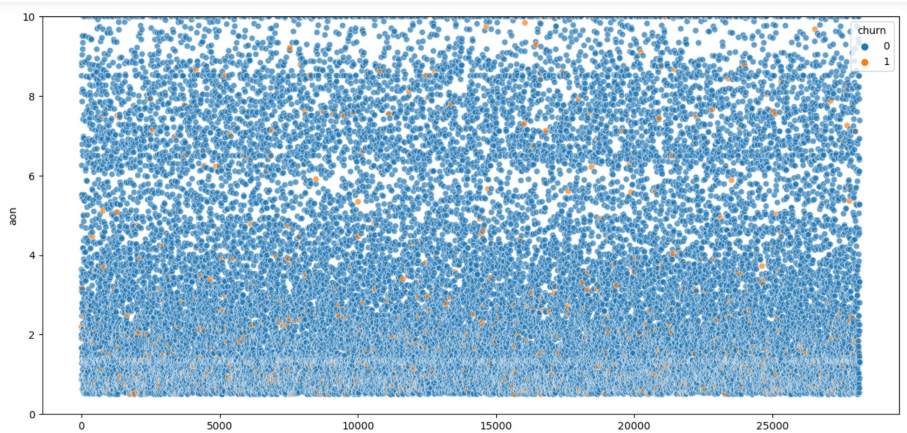
Data cleaning

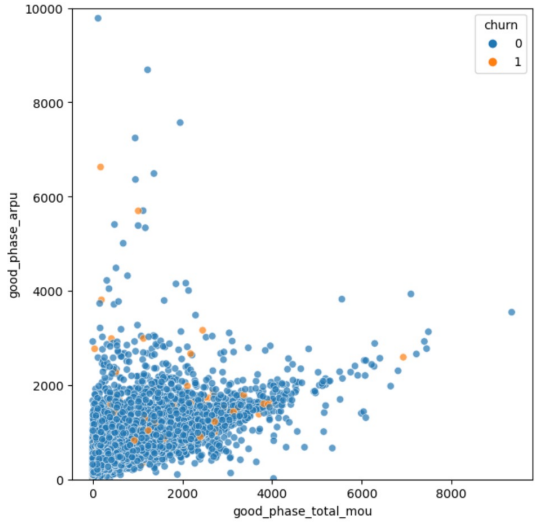
- Check and handle duplicate data.
- Check and handle NA values and missing values.
- Drop columns, if it contains large amount of missing values and not useful for the analysis.
- Imputation of the values, if necessary.
- Check and handle outliers in data.

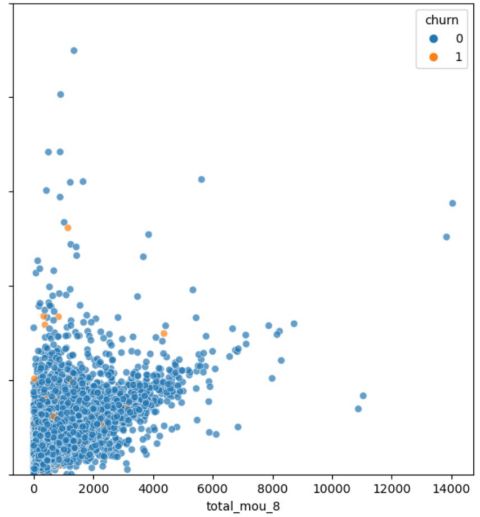
Exploratory data Analysis

- Univariate data analysis: value count, distribution of variable etc.
- Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Data preparation, Standardization, Handling Class Imbalance, Principal Component Analysis(PCA)
- Selecting the best classification model: Logistic regression, Decision Tree, Random Forest
- Validation of the best model.

UNIVARIATE AND MULTIVARIATE ANALYSIS







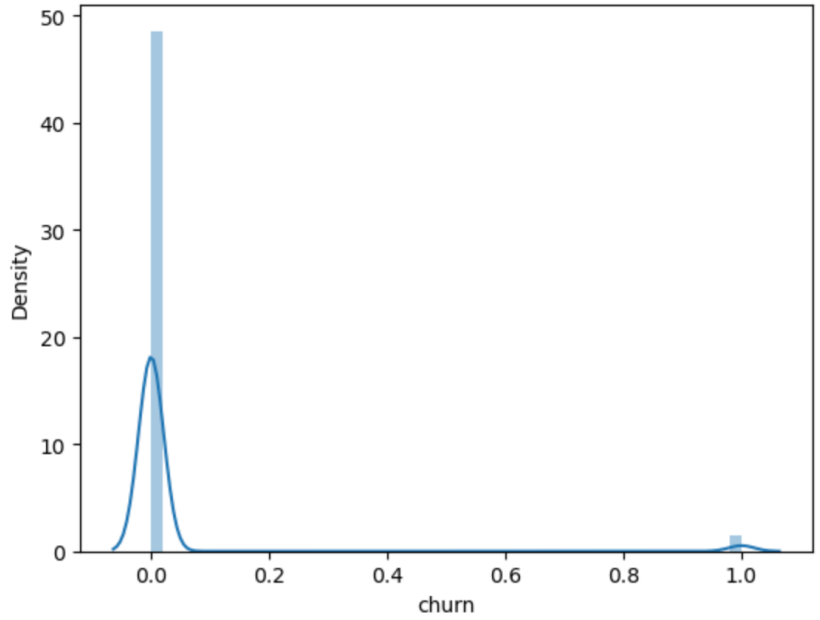
Observation

- It is clearly evident that MOU has dropped significantly for the churners in the action phase (8th month) thus dropping the revenue generated from them
- It is interesting to note that MOU is between 0-2000 which means that revenue is highest in that region which implies that these users has used other services that were boosting the revenue

HANDLING CLASS IMBALANCE

```
# Distribution of target variable

sns.distplot(churn_telecom_data['churn'])
plt.show()
```



Observation

- Though the target variable is not skewed, it is highliy imbalanced. The number of non-churners in the dataset is around 94%
- This imbalance will be handled using SMOTE algorithm

PRINCIPAL COMPONENT ANALYSIS

PCA

74]:
```
X.shape
```

74]:
```
(28163, 55)
```

75]:
```
from sklearn.decomposition import PCA

pca = PCA(n_components=25)
X_pca = pca.fit_transform(X_res)
X_pca.shape
```

75]:
```
(54736, 25)
```

MODEL BUILDING

- As the dependent variable is categorical hence the general model is classification model.
- Now classification taught are- Logistic Regression, Decision Tree and Random Forest.
- Hence, all three models have been made and tested on various parameters and results like accuracy, precision, ROC.
- After analysing all, the three models, the best model came out to be Random Forest.

CONCLUSION

- Given our business problem, to retain their customers, we need higher recall. As giving an offer to an user not going to churn will cost less as compared to loosing a customer and bring new customer, we need to have high rate of correctly identifying the true positives, hence recall.
- When we compare the models trained we can see the tuned random forest is performing the best, which is highest accuracy along with highest recall i.e. 95%. So, we will go with random forest.

FINAL MODEL/CONCLUSION

In [132]:
```
final_model = RandomForestClassifier(max_depth=30, min_samples_leaf=5, n_jobs=-1,
                        random_state=25)
```

In [133]:
```
y_train_pred = rf_best.predict(X_train)
y_test_pred = rf_best.predict(X_test)



# Print the report
print("Report on train data")
print(metrics.classification_report(y_train, y_train_pred))

print("Report on test data")
print(metrics.classification_report(y_test, y_test_pred))
```

```
Report on train data
              precision    recall  f1-score   support

           0       1.00      0.98      0.99     19110
           1       0.98      1.00      0.99     19205

    accuracy                           0.99     38315
   macro avg       0.99      0.99      0.99     38315
weighted avg       0.99      0.99      0.99     38315

Report on test data
              precision    recall  f1-score   support

           0       0.97      0.94      0.95      8258
           1       0.94      0.97      0.95      8163

    accuracy                           0.95     16421
   macro avg       0.95      0.95      0.95     16421
weighted avg       0.95      0.95      0.95     16421
```

STRATERGIES TO MANAGE CUSTOMER CHURN/ INSIGHTS

- Given our business problem, to retain the customers, we need higher recall. Since giving an offer to a customer who is not going to churn will cost less as compared to losing and to get a new customer, we need to have high rate of correctly identifying the true positives, hence recall.
- When we compare the trained models, we can see that tuned random forest is performing the best which is highest accuracy as well as highest recall i.e., 95%. So we will go with random forest.

CONCLUSIONS

Strategies to manage customer churn

The top 10 predictors are:

Feature
loc_og_mou_8
total_rech_num_8
monthly_3g_8
monthly_2g_8
good_phase_loc_og_mou
good_phase_total_rech_num
last_day_rch_amt_8
std_ic_t2t_mou_8
sachet_2g_8
aon

- We can see most of the top predictors are from action phase as drop in engagement in prominent in that phase

Some of the factors we noticed while performing EDA which can be clubed with these insights are:

- Users whose maximum recharge amount is less than 200 even in the good pahse, should have a tag and re-evaluated time to time as they are more likely to churn
- Users that have been with the network for less than 4 years, should be monitored time to time, as from data we can see that users who have been associated with the network for less than 4 years tend to churn more
- MOU is one of the major factors, but data especially VBC if the user is not using a data pack if another factor to look out