Telecom Churn Case Study

By Manisha, Deepak & Saurabh

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 analyses 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

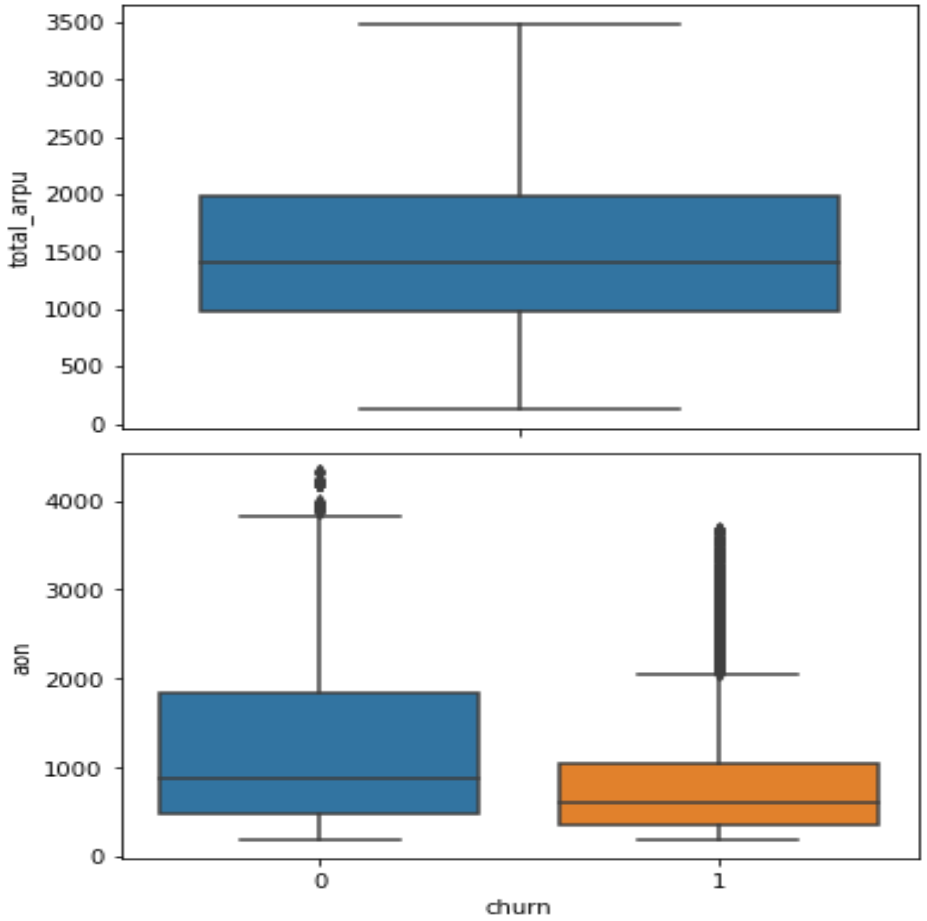
Business Objective

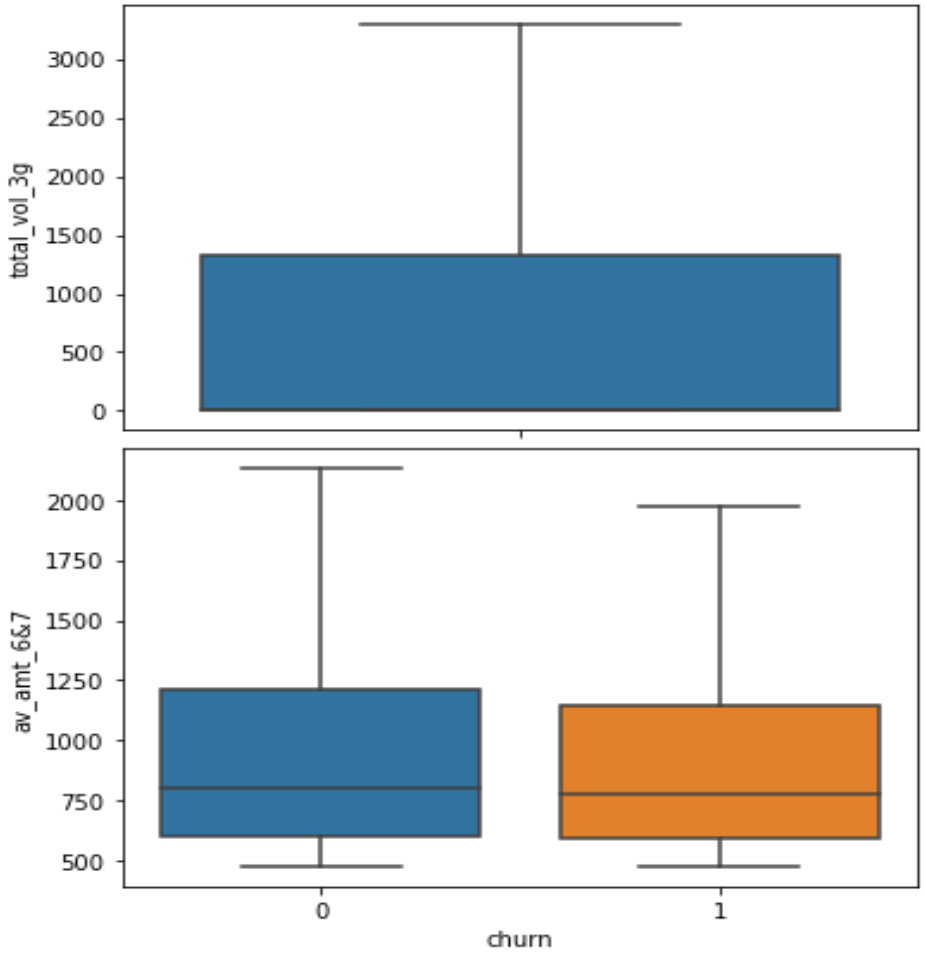
The business objective is to predict the churn in the last (i.e. the ninth) month using the data (features) from the first three months. To do this task well, understanding the typical customer behavior during churn will be helpful.

Approach

- Reading and understanding the data.
- Handling missing values and further cleaning the data.
- Deriving new features.
- Filtering high value customers.
- Tagging churners and removing attributes of the churn phase.
- Univariate and Bivariate analysis.
- Building the models.
- Conclusions and recommendations.

EDA- Telecom Churn Analysis

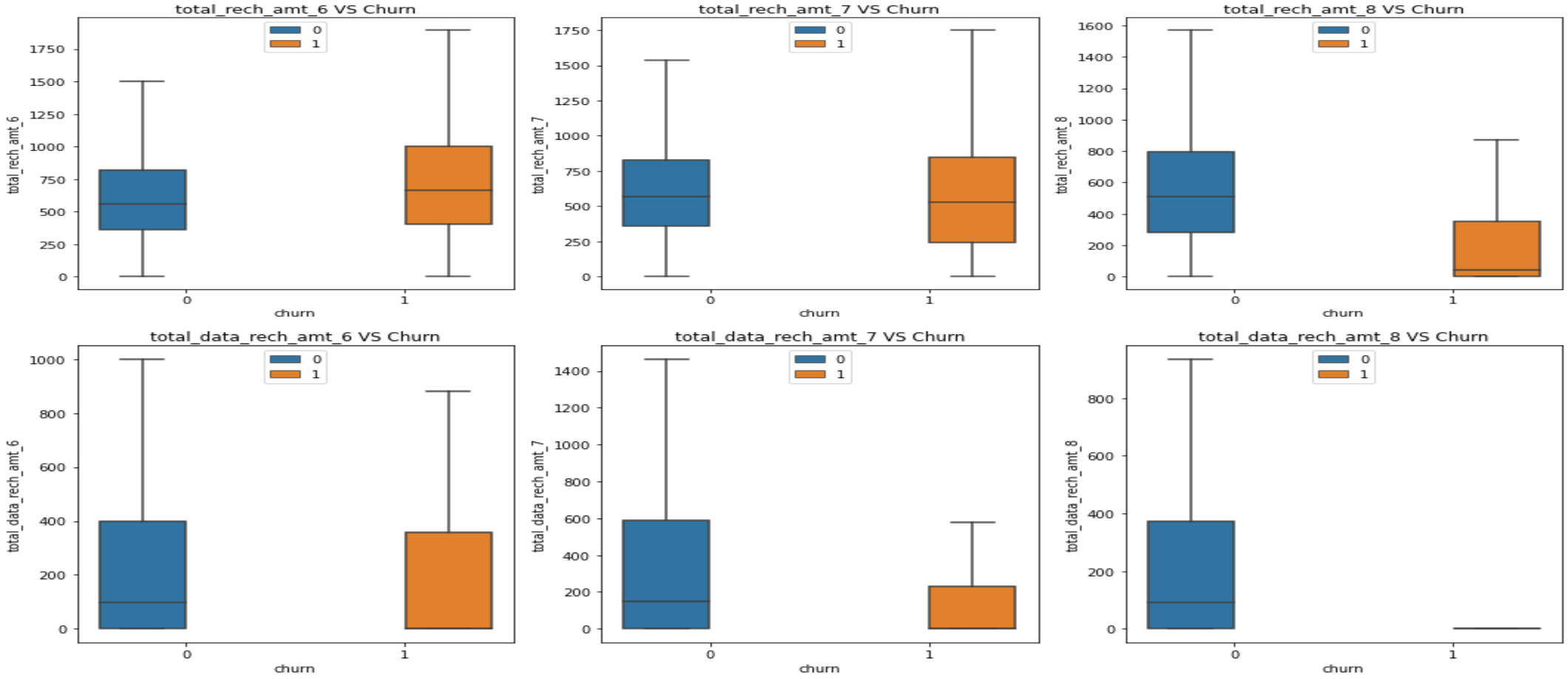




Comments

- The average of total revenue generated per user is 1627.87. Median of total_arpu is 1409.3
- Around 50% of the customers do not use the 3g services, because of which the median value is 0. Average of total 3g volume used by customers is 1065.9
- Churners have slightly lower median and 75th percentile for av_amt_6&7 compared to non churners
- The 75th percentile of age on network of churned customers is 1039 days, which means is that among the people who have churned 75% of the them have left within first 3 yrs

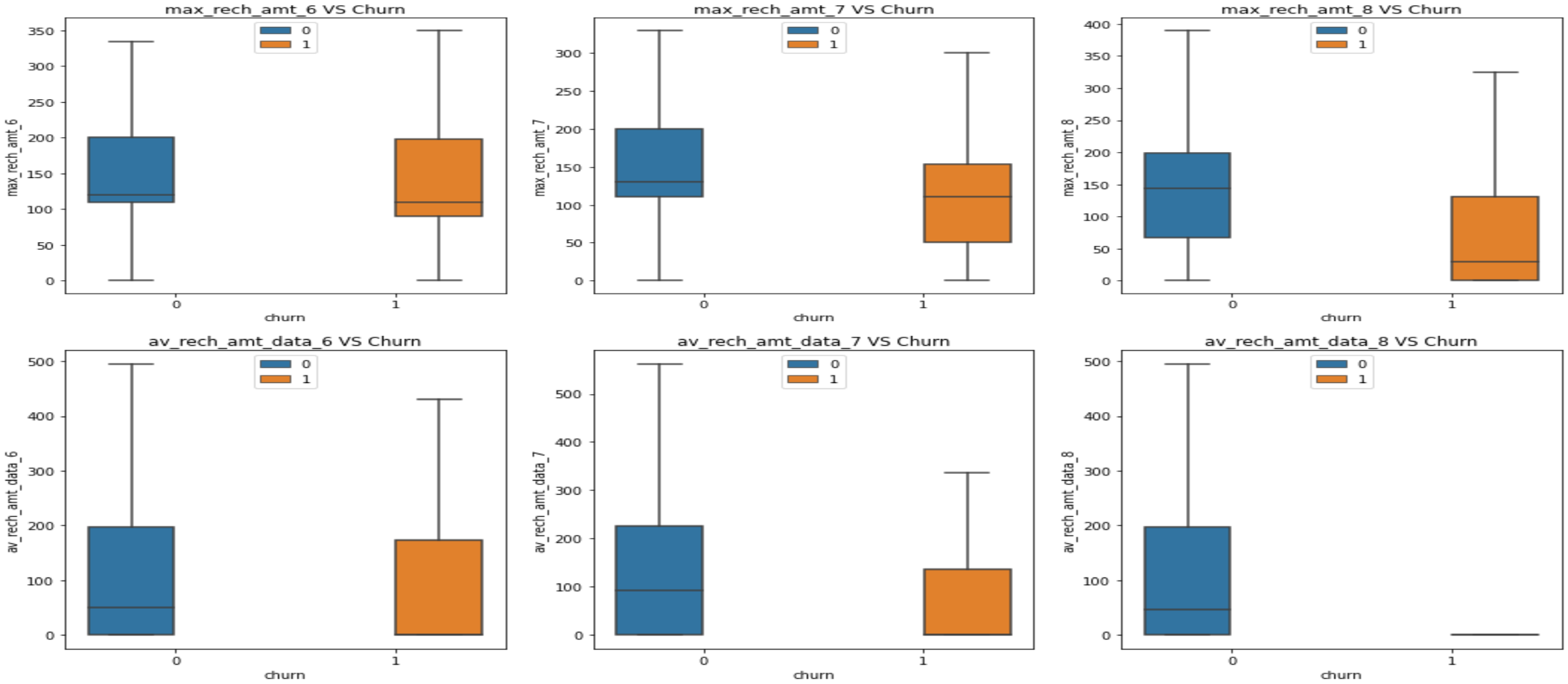
EDA- Telecom Churn Analysis



Comments

- There is a drop in the total amount re-charged by churned customers in the 8th month(action phase)
- The median of total amount re-charged by customers who have not churned remains nearly same for all the months
- Churned customers have made nearly zero data recharge in the 8th month
- 75th percentile and maximum values of total_data_rech_amt decreases over 6-8 month period

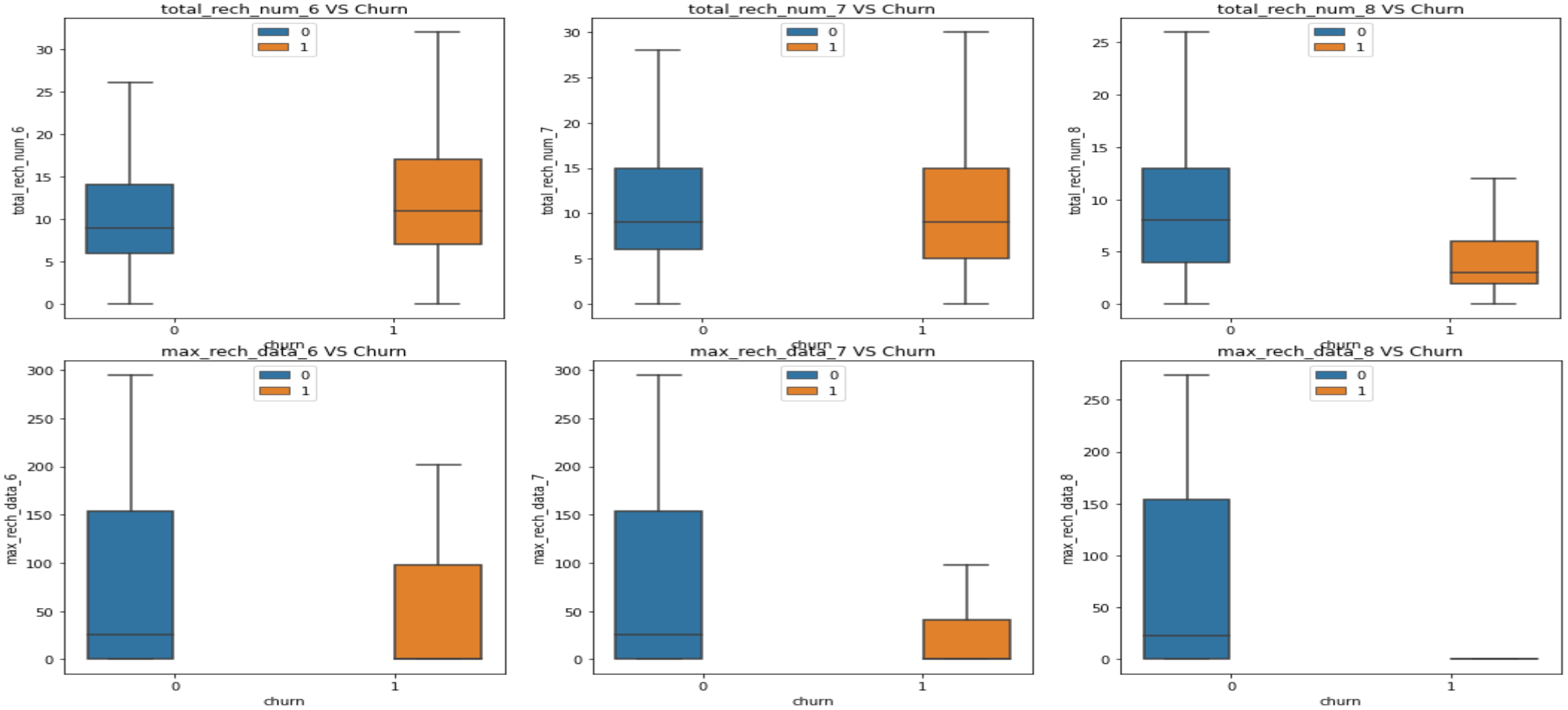
EDA- Telecom Churn Analysis



Comments

- There is a drop in 25th percentile, median and 75th percentile of maximum recharge amount made by churned customers in 8th month
- The average recharge amount(for data) paid by Churned customers is nearly zero in the 8th month
- 75th percentile and maximum values of av_rech_amt_data decreases over 6-8 month period

EDA- Telecom Churn Analysis

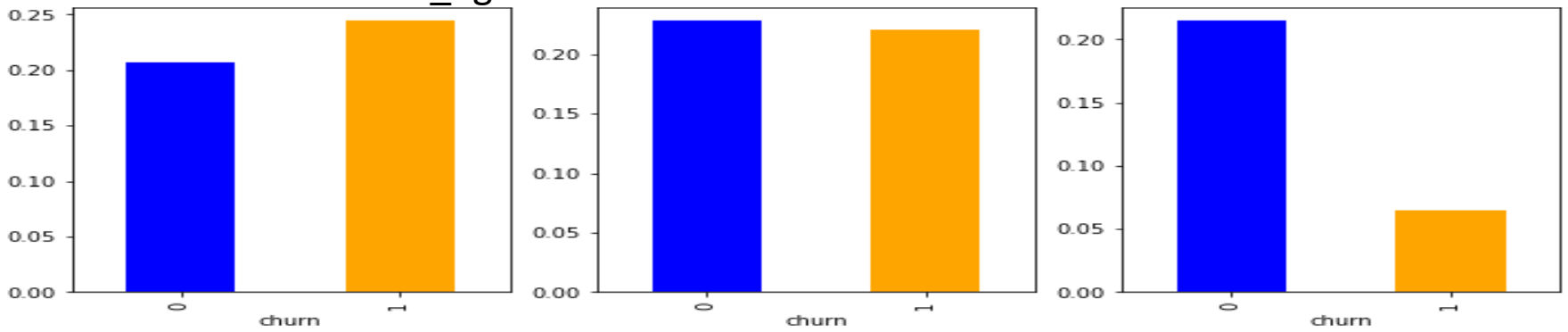


Comments

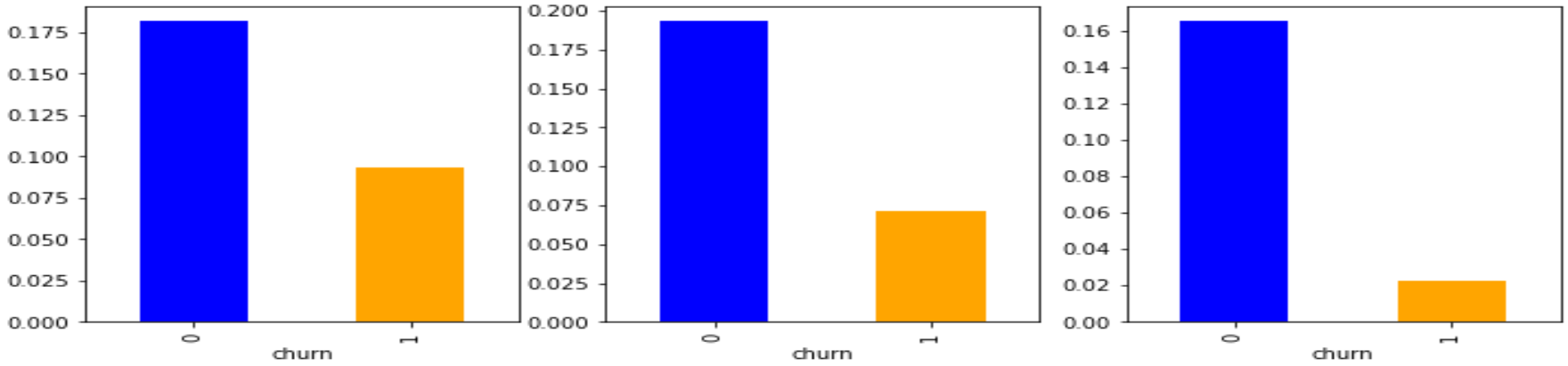
- The total number of recharge done by churned customers were higher than the customer not churned in the 6th month.
- However during 7th month number of recharge done by churned customers got reduced and became nearly equal to that of non churned customers
- In the action phase this number for churned customers got reduced drastically.
- There is a drop in maximum recharge done for data in 8th month by churned customers

EDA- Telecom Churn Analysis

Mean of sachet_3g for churners and non churners in different months



Mean of monthly_2g for churners and non churners in different months



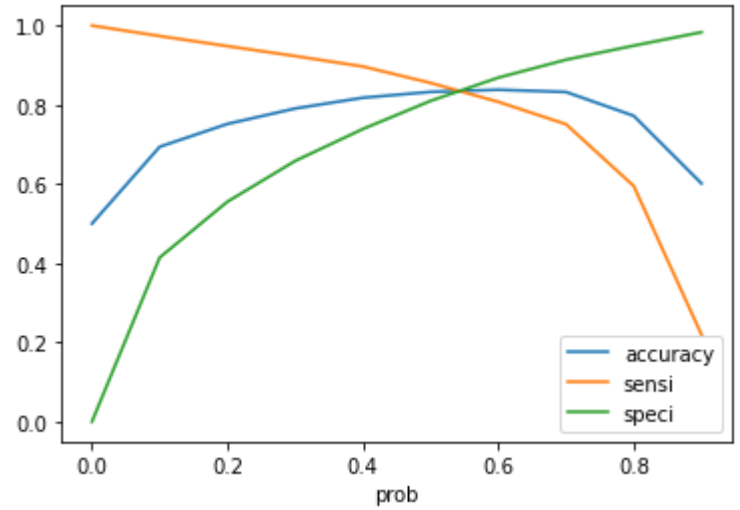
Comments

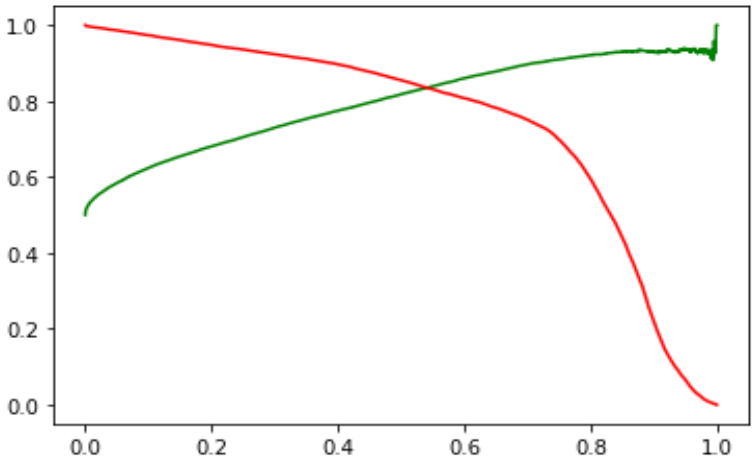
- The mean sachet_3g of churners has dropped drastically in the 8th month
- Churners have smaller mean value for monthly_2g in all three months
- mean monthly_2g for churners decreases over months

Model Building Approach

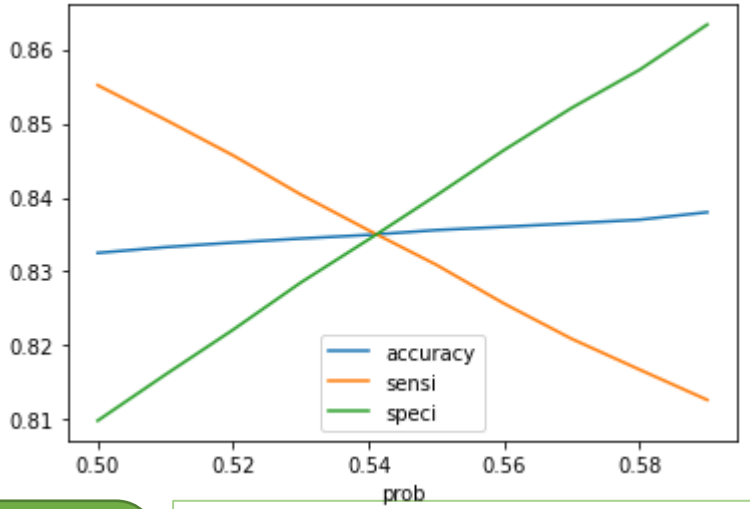
- **Building the models**
  - Train and test split
  - Feature scaling
  - first training model
  - Logistic regression model using REF selected features
  - Multiple logistic regression models
  - PCA
  - Logistic model with principal components
  - Random forest model
  - Random forest model with tuned hyperparameters

Model Evaluation (Train set)





Optimum cut off seems to be between 0.5 and 0.6 from precision- recall view

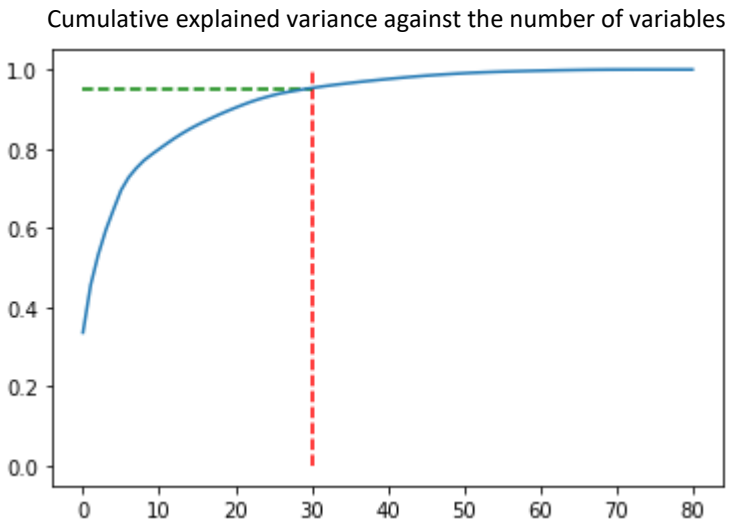


Comments

- Earlier with 0.5 cut-off
  - accuracy: 83.25%
  - sensitivity : 85.52%
  - specificity : 80.99%

•After choosing the optimum cut-off at 0.54
- accuracy : 83.49%
- sensitivity : 83.56%
- specificity : 83.43%

PCA



Random FOREST with PCA

Model evaluation on test set

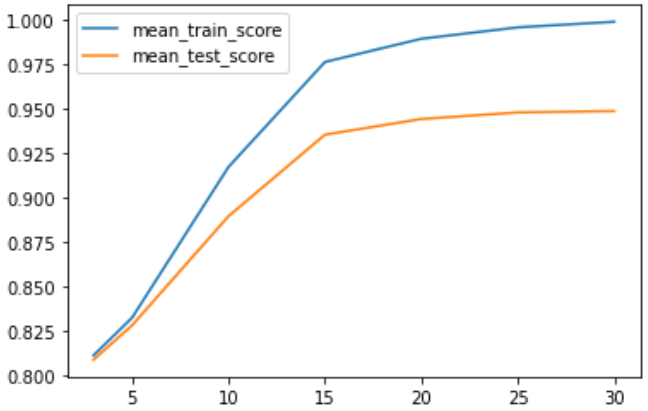
- Accuracy_score: 91%
- Recall_score: 55%
- Precision_score: 47%
- Auc_score: 75%
- F1_score: 51%

Comments

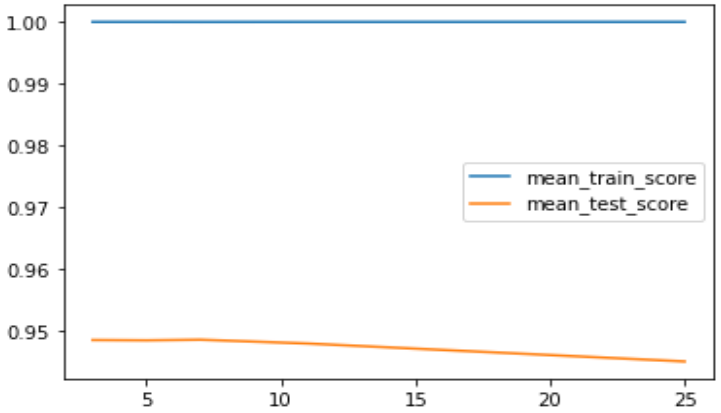
- For further analysis, let's select X_tr_pca and y_tr as our new training sets, as we are planning to go ahead with 30 feature variables which would explain 95% of the variance in data and a smote performed response variable.
- For testing models, let's use x_test_pca with selected 30 components.

Hyper parameter tuning (for train and test sets)

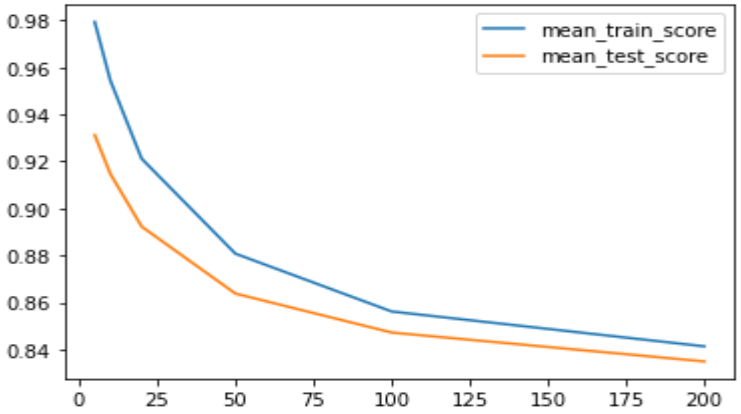
max_depth vs accuracy scores



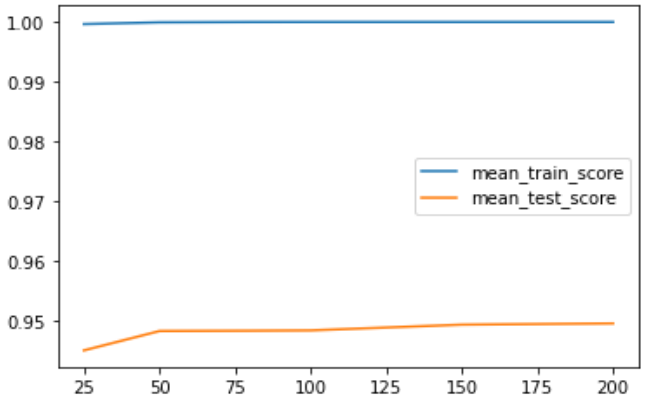
max_features vs accuracy scores



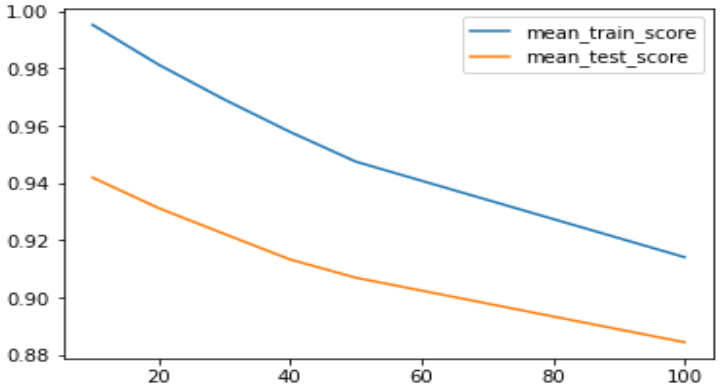
min_samples_leaf vs accuracy scores



n_estimators vs accuracy scores



min_samples_split vs accuracy scores



Final Random forest model

Model evaluation on train set

- Accuracy_score: 98%
- Recall_score: 99%
- Precision_score: 96%
- Auc_score: 98%
- F1_score: 98%

Model evaluation on test set

- Accuracy_score: 90%
- Recall_score: 62%
- Precision_score: 43%
- Auc_score: 77%
- F1_score: 51%

Comments

- As the value of max_depth increases the accuracy of model increases on both train and test sets till a point near 15 and then there is not much change in accuracy with increase in max_depth
- There is not much change in accuray of model on training set with different values of max_features But the test score decreases after some point around 7
- Model accuracy reduces with increase in min_samples_leaf
- Test accuracy is very stable after n_estimators = 50. So we use n_estimator as 50 or 80 while building the model
- accuracy is high at min_samples_split= 10. we consider 10-20 while fine tuning.

Conclusions and recommendations

- Among all the different types of models we made built, Random forest models seem to be the the better at predicting telecom churn
- Random RF model produces an accuracy of 0.91 in default overfit model and 0.9 on rf model with tuned hyper parameters.
- The 75th percentile of age on network of churned customers is 1039 days, which means is that among the people who have churned 75% of the them have left within first 3 yrs.
- The first three years are crucial for the retention of customers, so the operator needs to provide many offers during this period.
- The average of average revenue generated per user in 6th and 7th months of churners were very high compared to non churners. Having an higher arpu_avg_6_7 means that the customer is a heavy user of the services provided. Since these individuals are paying alot , they are more likely to look for other services providers with cheaper prices/better offers. So in order to reduce churning the company needs to satisfy the people with higher 'average revenue per user'.
- The telecom company should focus on total amount recharged in eighth month as it can be a good indicator of churning
- So if the total amount recharged in eighth month is low, then there is very high chance that the customer will churn
- Among the churned customers the 2g/3g usage volume was very low, this could be because the signal is not good enough in certain places. So the company needs to provide better 3g/3g network coverage in those areas.