Telecom Churn Case Study

Shashank A
Hritik Vijay Thorat
Aiman Maryam Zakriya
(DS160)

Problem Statement

- To minimize customer churn, telecom companies must forecast which customers are most likely to leave.
- In this project, we will examine customer-level data from a prominent telecom firm, develop predictive models to identify customers at a high risk of churn, and pinpoint the key factors contributing to churn.
- The primary objective is to retain our most profitable customers. Since it is known that the cost of losing a customer is higher than gaining one.

Steps Followed

- Reading and Understanding the data
- Data Processing
  - Handling missing values
  - Identifying high value customers
  - Tagging churners
  - Outlier treatment
  - Adding new features
- EDA
  - Univariate Analysis
  - Bivariate Analysis

Steps Followed (Contd.)

- Test-Train Split
- Dealing with Data Imbalance
- Feature Scaling
- Modelling with PCA
  - Logistic Regression
  - Support Vector Machine(SVM)
  - Decision Tree
  - Random Forest
- Modelling without PCA
  - Logistic Regression
- Summary
- Recomendations

Reading and Understanding the data

- The dataset was imported and read using basic python commands such as shape, info and describe.
- It was clear that the dataset had 99999 rows and 226 columns.

Data Processing – Handling missing values

- The data was checked for missing value percentage in all columns and columns with more than 30% was dropped.
- Date columns were dropped as the it was not required in our analysis.
- circle_id column is dropped as this column has only one unique value. Hence there will be no impact of this column on the data analysis.
- Similarly, we have handled missing values for which MOU are null for respective months.

Data Processing – Identifying high value customers

- Creating column `avg_rech_amt_6_7` by summing up total recharge amount of month 6 and 7. Then taking the average of the sum.
- Then Considering customers who have recharged more than or equal to 70^{th} percentile.

Data Processing – Tagging churners

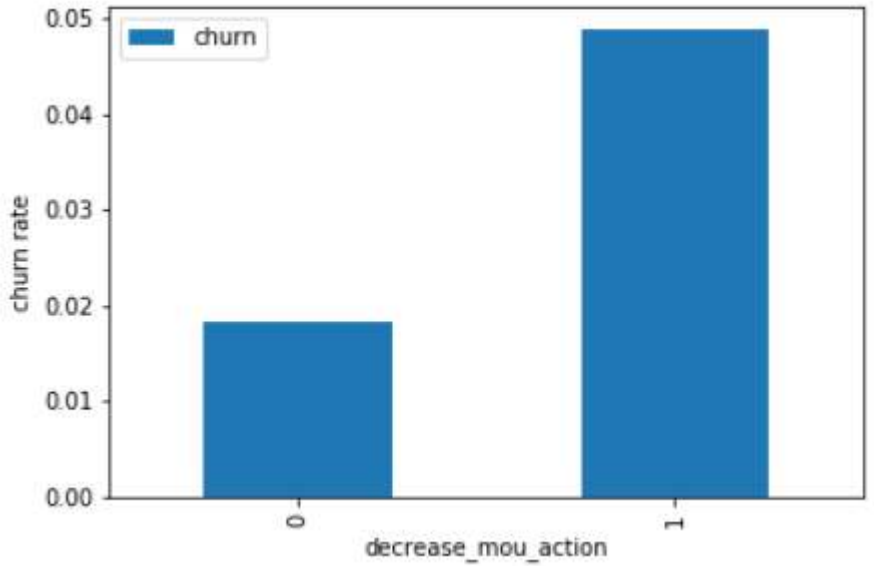
- Tagging the churned customers (churn=1, else 0) based on the fourth month as follows:
  - Those who have not made any calls (either incoming or outgoing) and
  - Those who have not used mobile internet even once in the churn phase.
- After tagging churners, we need to remove all the attributes corresponding to the churn phase (all attributes having ‘ _9’, etc. in their names).

Data Processing – Outlier treatment & Deriving new features

- On the processed data, outlier treatment was performed.
- New features were added for better model building. All the features are mentioned with description on the code.

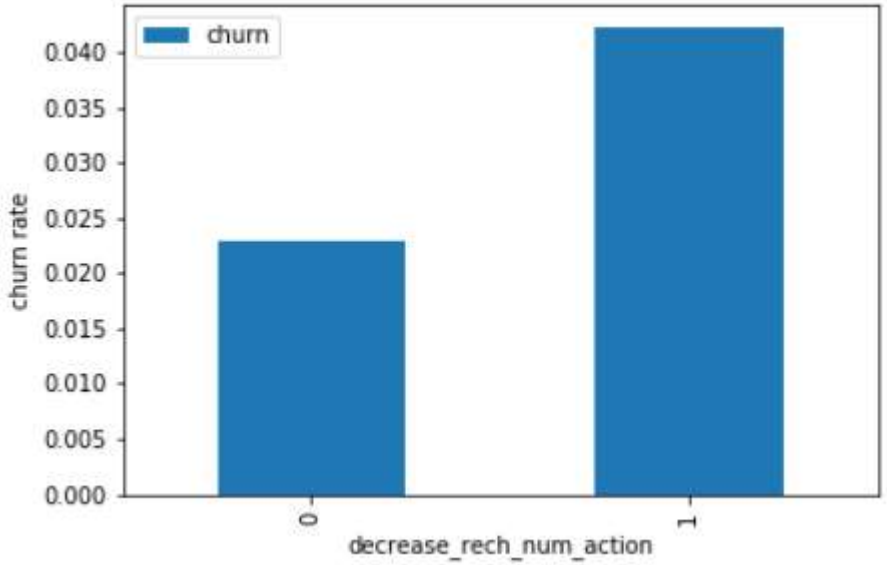
EDA – Univariate Analysis

- Churn rate on the basis whether the customer decreased her/his MOU in action month.
- We can see that the churn rate is more for the customers, whose minutes of usage(mou) decreased in the action phase than the good phase.



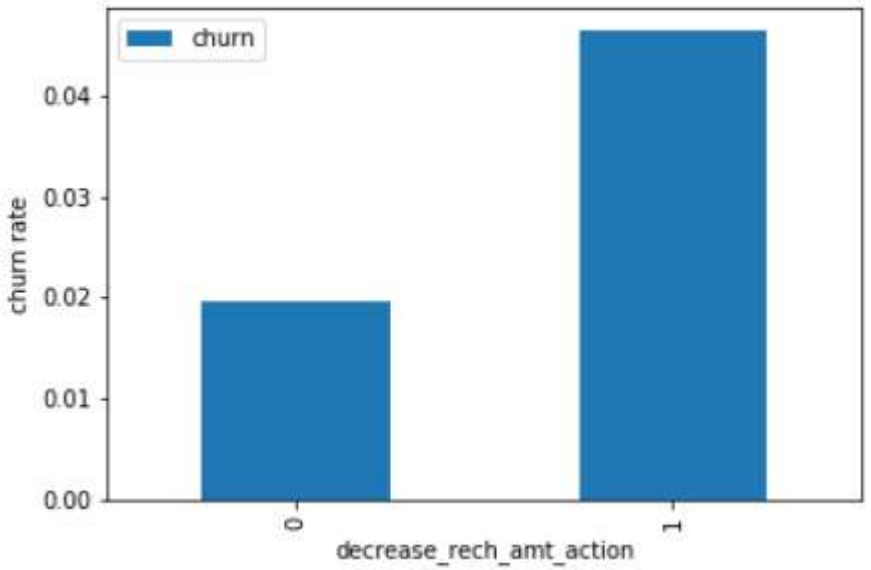
EDA – Univariate Analysis

- Churn rate on the basis whether the customer decreased her/his number of recharge in action month.
- As expected, the churn rate is more for the customers, whose number of recharge in the action phase is lesser than the number in good phase.



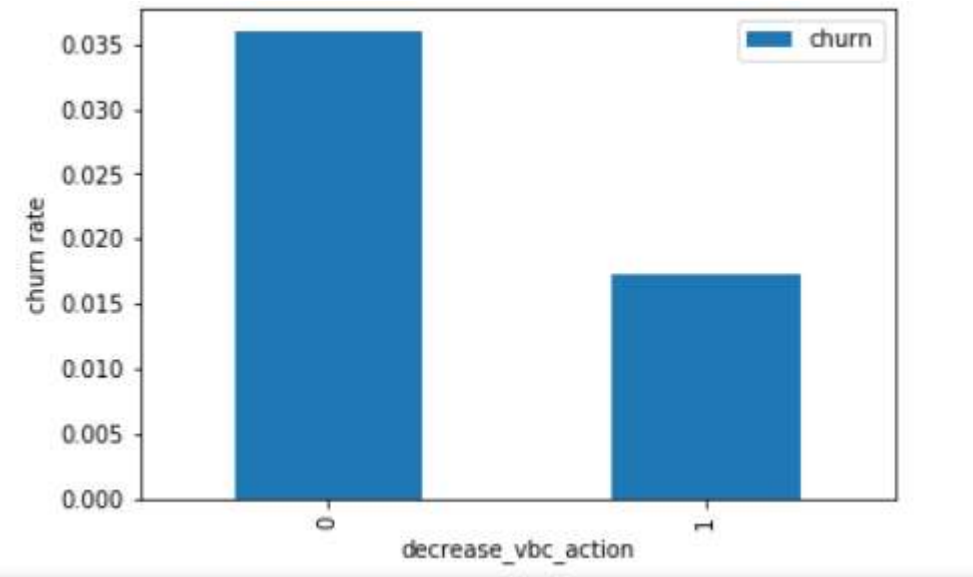
EDA – Univariate Analysis

- Churn rate on the basis whether the customer decreased her/his amount of recharge in action month.
- Here also we see the same behaviour. The churn rate is more for the customers, whose amount of recharge in the action phase is lesser than the amount in good phase.



EDA – Univariate Analysis

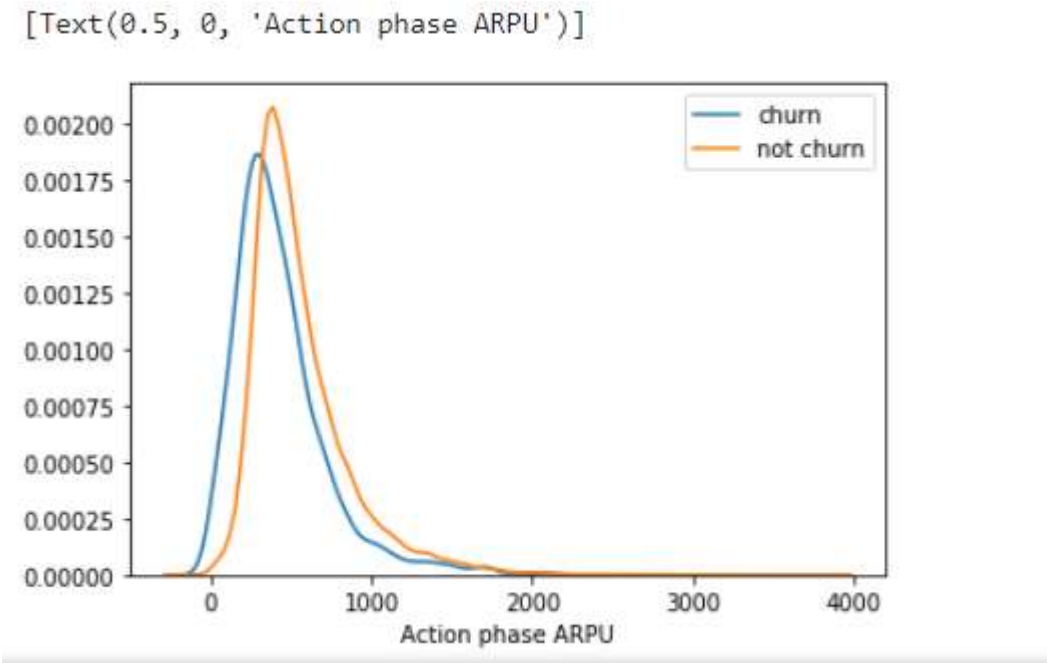
- Churn rate on the basis whether the customer decreased her/his volume-based cost in action month.
- Here we see the expected result. The churn rate is more for the customers, whose volume-based cost in action month is increased. That means the customers do not do the monthly recharge more when they are in the action phase.



EDA – Univariate Analysis

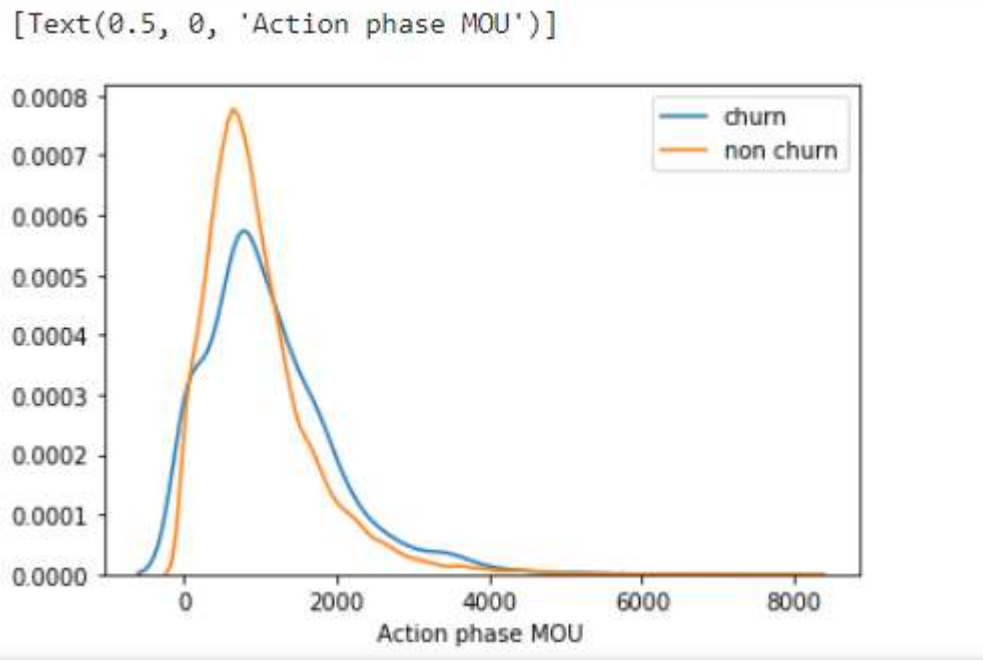
- Analysis of the average revenue per customer (churn and not churn) in the action phase.
- Average revenue per user (ARPU) for the churned customers is mostly densed on the 0 to 900. The higher ARPU customers are less likely to be churned.

- ARPU for the not churned customers is mostly densed on the 0 to 1000



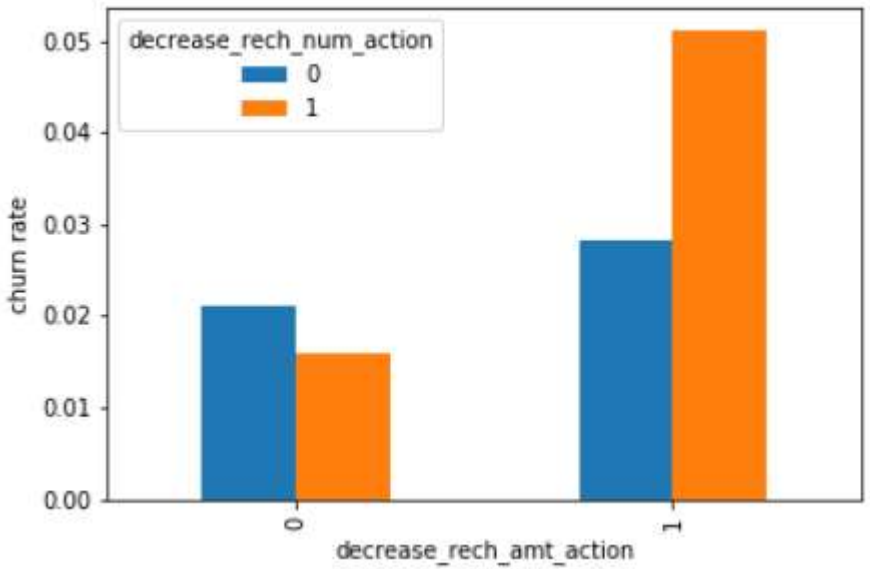
EDA – Univariate Analysis

- Analysis of the minutes of usage MOU (churn and not churn) in the action phase.
- Minutes of usage(MOU) of the churn customers is mostly populated on the 0 to 2500 range. Higher the MOU, lesser the churn probability.



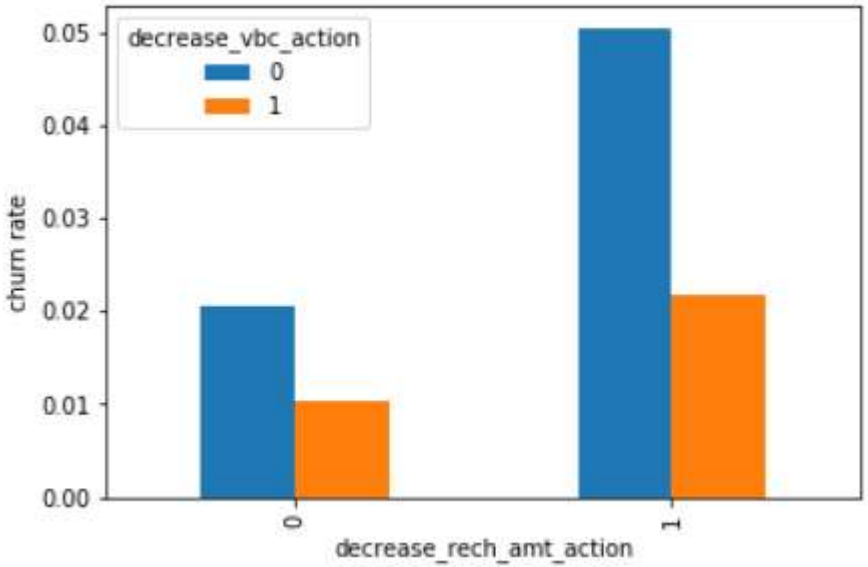
EDA – Bivariate Analysis

- Analysis of churn rate by the decreasing recharge amount and number of recharge in the action phase.
- We can see from the above plot, that the churn rate is more for the customers, whose recharge amount as well as number of recharge have decreased in the action phase than the good phase.



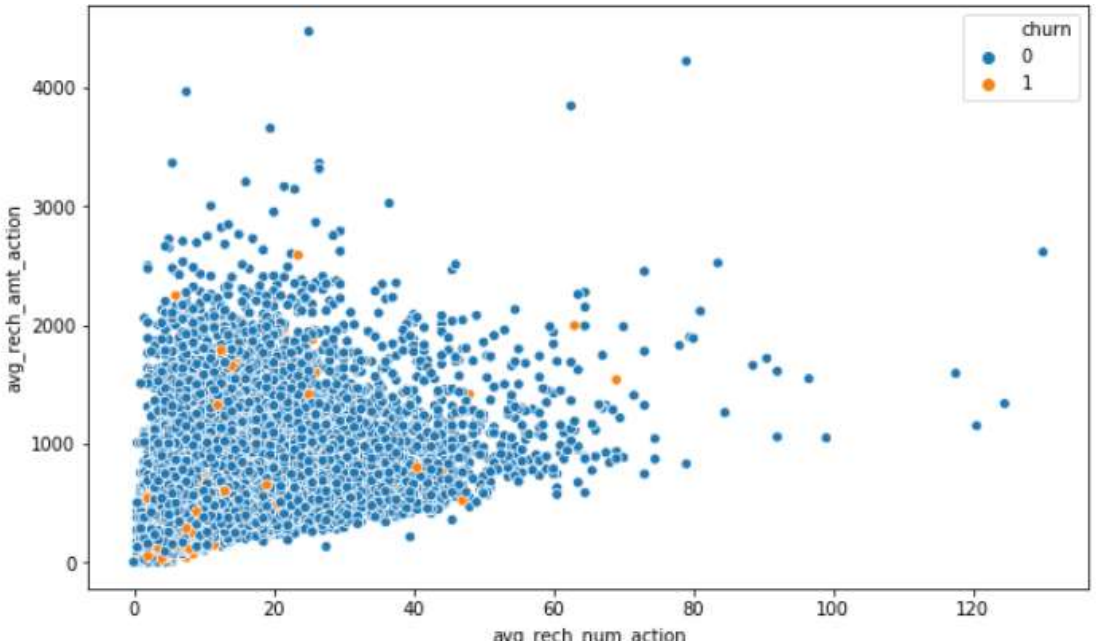
EDA – Bivariate Analysis

- Analysis of churn rate by the decreasing recharge amount and volume-based cost in the action phase.
- Here, also we can see that the churn rate is more for the customers, whose recharge amount is decreased along with the volume-based cost is increased in the action month.



EDA – Bivariate Analysis

- Analysis of recharge amount and number of recharge in action month.
- We can see from the above pattern that the recharge number and the recharge amount are mostly propotional. More the number of recharge, more the amount of the recharge.



Model Building

- Test – Train Split was performed; Class imbalance was dealt with and feature scaling was done.
- The next step is to build models

Model with PCA

- We observed that 60 components explain most more than 90% variance of the data. So, we will perform PCA with 60 components.
- We are more focused on higher Sensitivity/Recall score than the accuracy.
- Because we need to care more about churn cases than the not churn cases. The main goal is to retain the customers, who have the possibility to churn. There should not be a problem, if we consider few not churn customers as churn customers and provide them some incentives for retaining them. Hence, the sensitivity score is more important here.

Logistic Regression with PCA

Model summary

- Train set
  - Accuracy = 0.86
  - Sensitivity = 0.89
  - Specificity = 0.83
- Test set
  - Accuracy = 0.83
  - Sensitivity = 0.81
  - Specificity = 0.83

Overall, the model is performing well in the test set, what it had learnt from the train set.

Support Vector Machine(SVM) with PCA

Model summary

- Train set
  - Accuracy = 0.89
  - Sensitivity = 0.92
  - Specificity = 0.85
- Test set
  - Accuracy = 0.85
  - Sensitivity = 0.81
  - Specificity = 0.85

Decision tree with PCA

- We saw from the model performance that the Sensitivity has been decreased while evaluating the model on the test set. However, the accuracy and specificity is quite good in the test set.

Model summary

- Train set
  - Accuracy = 0.90
  - Sensitivity = 0.91
  - Specificity = 0.88
- Test set
  - Accuracy = 0.86
  - Sensitivity = 0.70
  - Specificity = 0.87

Random forest with PCA

- We can see from the model performance that the Sensitivity has been decreased while evaluating the model on the test set. However, the accuracy and specificity is quite good in the test set.

Model summary

- Train set
  - Accuracy = 0.84
  - Sensitivity = 0.88
  - Specificity = 0.80
- Test set
  - Accuracy = 0.80
  - Sensitivity = 0.75
  - Specificity = 0.80

Final conclusion with PCA

- After trying several models, we can see that for achieving the best sensitivity, which was our ultimate goal, the classic Logistic regression or the SVM models preforms well. For both the models the sensitivity was approx. 81%. Also, we have good accuracy of approx. 85%.

Final conclusion with PCA

- After trying several models, we can see that for achieving the best sensitivity, which was our ultimate goal, the classic Logistic regression or the SVM models preforms well. For both the models the sensitivity was approx. 81%. Also, we have good accuracy of approx. 85%.

Logistic regression with No PCA

- We can see that the logistic model with no PCA has good sensitivity and accuracy, comparable to the models with PCA. So, we can go for the more simplistic model such as logistic regression with PCA as it explains the important predictor variables as well as the significance of each variable. The model also helps us to identify the variables that should be acted upon for deciding on churned customers. Hence, the model is more relevant in terms of explaining to the business.

Model summary

- Train set
  - Accuracy = 0.84
  - Sensitivity = 0.81
  - Specificity = 0.83
- Test set
  - Accuracy = 0.78
  - Sensitivity = 0.82
  - Specificity = 0.78

Overall, the model is performing well in the test set, what it had learnt from the train set.

Recommendations

- Focus on customers whose usage of incoming local calls and outgoing ISD calls is low during the action phase, particularly in August.
- Identify customers with reduced outgoing charges in July and incoming charges in August.
- Customers whose value-based costs have increased during the action phase are more likely to churn compared to others, making them prime candidates for targeted offers.
- Customers with higher monthly 3G recharges in August are at a greater risk of churning.
- Customers showing a decline in STD incoming minutes for calls from operator T to fixed lines of T in August are more likely to leave.
- Customers with decreasing monthly 2G usage in August are also at a higher risk of churning.
- Customers experiencing a drop in incoming minutes for calls from operator T to fixed lines of T in August are more likely to churn.8)The variable roam_og_mou_8 has a positive coefficient of 0.7135, indicating that customers with increasing roaming outgoing minutes are more likely to churn.