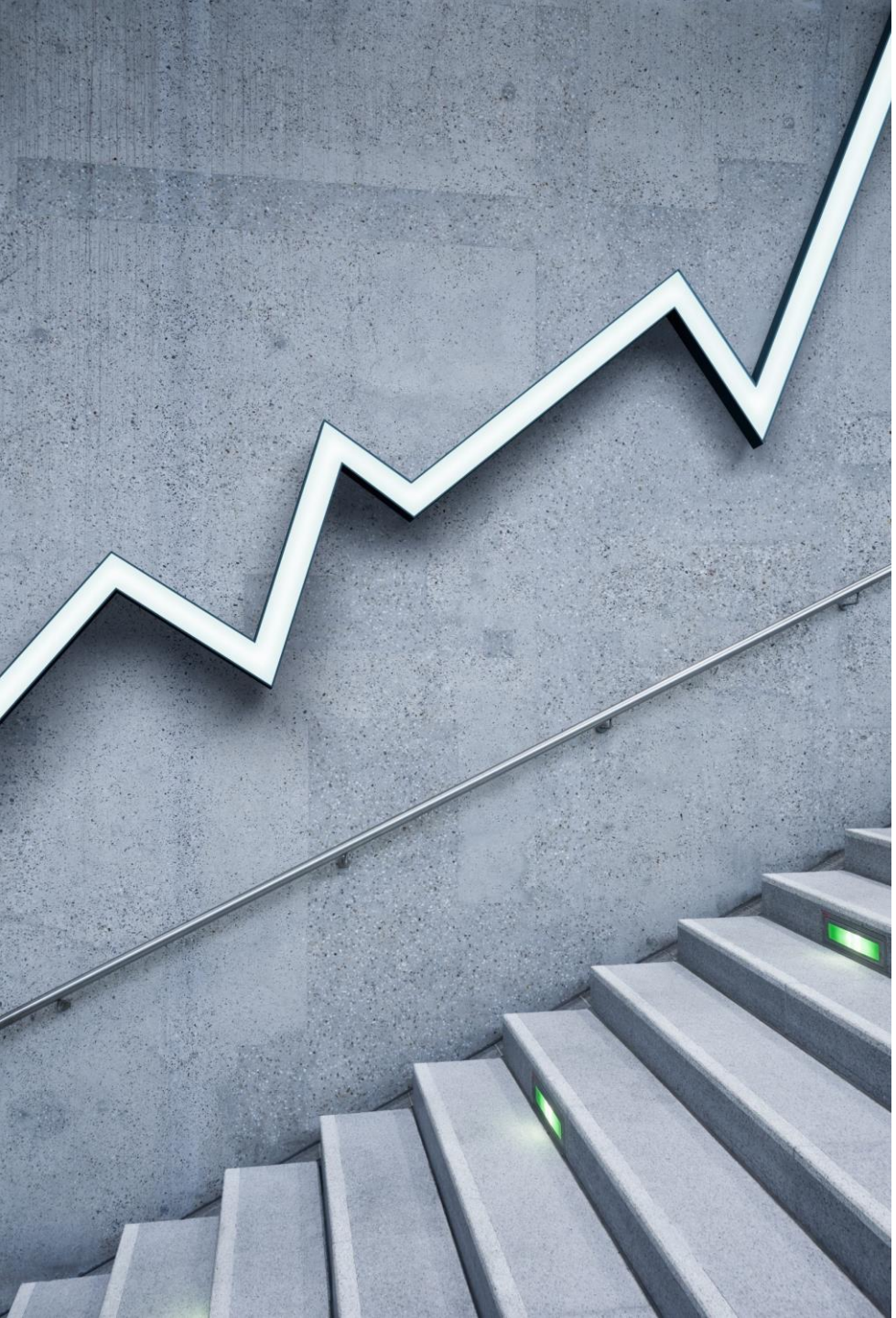


Telecom Churn

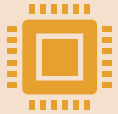
CASE STUDY

ABHILASH SIDDARAMAREDDY

ANISH LAKHOTIYA

ASHWANI SAINI

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.

To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.

GOAL OF THE CASE STUDY:

Retaining high profitable customers is the number one business goal

To analyze customer-level data of the telecom firm, build predictive models to identify customers at high risk of churn and identify the main indicators of churn.

Business Problems to Address:

- To predict high-value customer will churn or not.
- To identify important variables that are strong predictors of churn
- Build model with the main objective of identifying important predictor attributes which help the business understand indicators of churn.
- Recommend strategies to manage customer churn

STEP 1: IMPORTING LIBRARIES AND DATA

The following Python libraries were imported:

DATA ANALYSIS	DATA VISUALIZATION	MACHINE LEARNING
Numpy	Matplotlib	Statsmodels
Pandas	Seaborn	Scikit Learn

The datafile “telecom_churn_data.csv” was uploaded to start the EDA process. Upon uploading the dataframe was stored as “df”. The dataframe was inspected using head() function. The data frame contains 226 columns.

STEP 2: INSPECTING THE DATAFRAME

The shape of the Dataframe (99999 rows, 226 columns) was examined using the shape attribute.

To understand the data type of each column and number of missing values in each column, info() function was used.

The data type of 179 columns

- 179 columns is “float64” type
- 35 columns are “int64” type
- 12 columns are of “object” type

Step 3: Data Preparation

- Identifying the list of columns have more than 30% value missing and then dropping those columns.
- Next set of columns to be dropped are “date” as not required for analysis and “circle_id” has one unique value which will not have any impact. After dropping this columns left are 196.
- Find 70th percentile of the average recharge for 6^{th} & 7^{th} month.
- Filter the customers, who have recharged more than or equal to X(70^{th} percentile).
- We have 30001 rows and 196 columns
- Dropping rows having more than 50% value missing.
- Deleting the records for which MOU for Sep(9), Jun(6),Aug(8) & July(7) are null.
- We can see that we have lost almost 7% records. But we have enough number of records to do our analysis.

Step 3: Data Preparation (Continued)

1

Identify and tag customers (churn=1, else 0) for fourth month (Sep) who have not made any calls (either incoming or outgoing) AND have not used mobile internet even once in the churn phase.

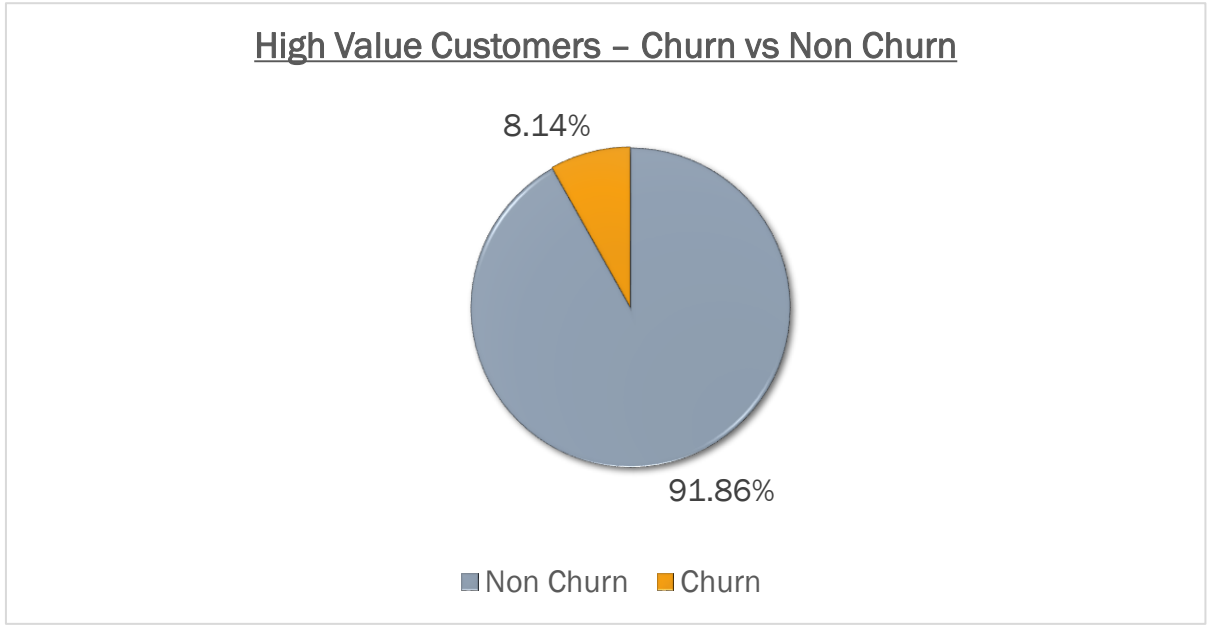
2

Check percentage after deleting the attributes for fourth month(Sep)

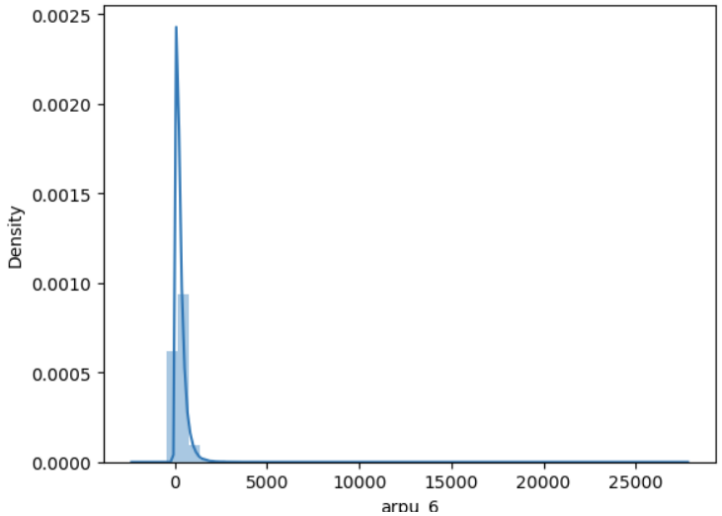
3

Derive new columns for below decreased in the action phase than the good phase

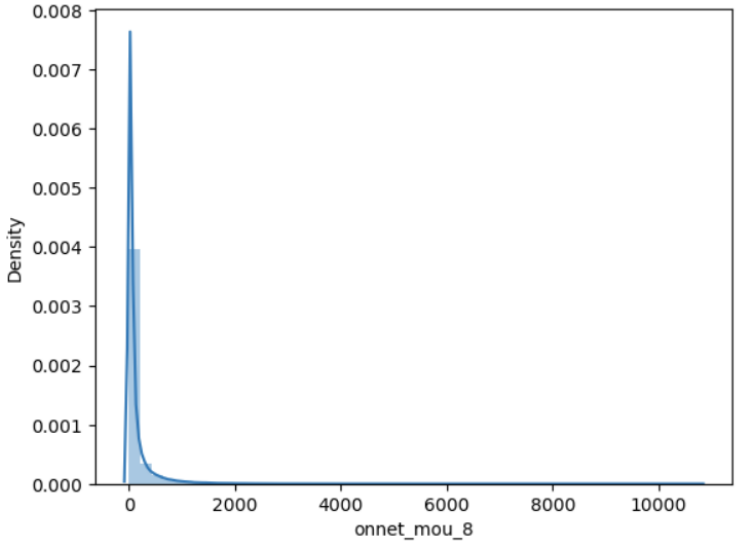
- Usage
- Recharge
- Average Revenue
- Volume based cost



Step 4: EDA - Univariate analysis

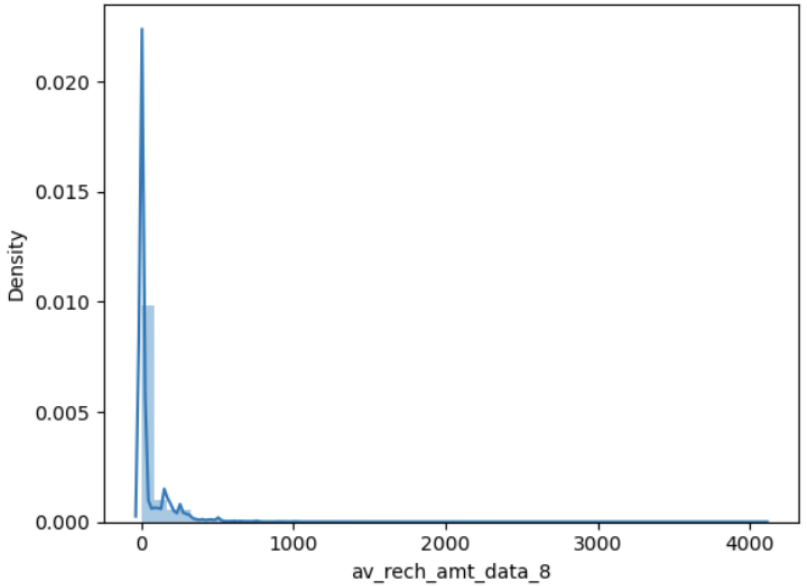


- Average revenue per user (ARPU) : Higher ARPU customers are less likely to be churned.



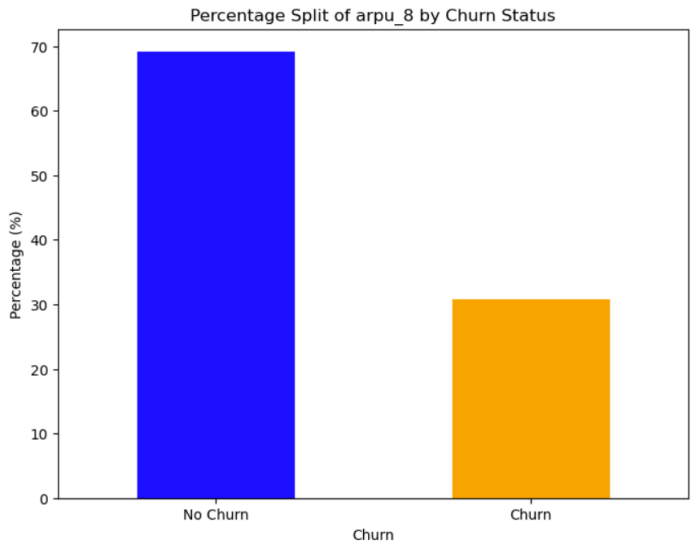
- Minutes of usage (MOU) : Higher the MOU, lesser the churn probability.

Step 4: EDA - Univariate analysis

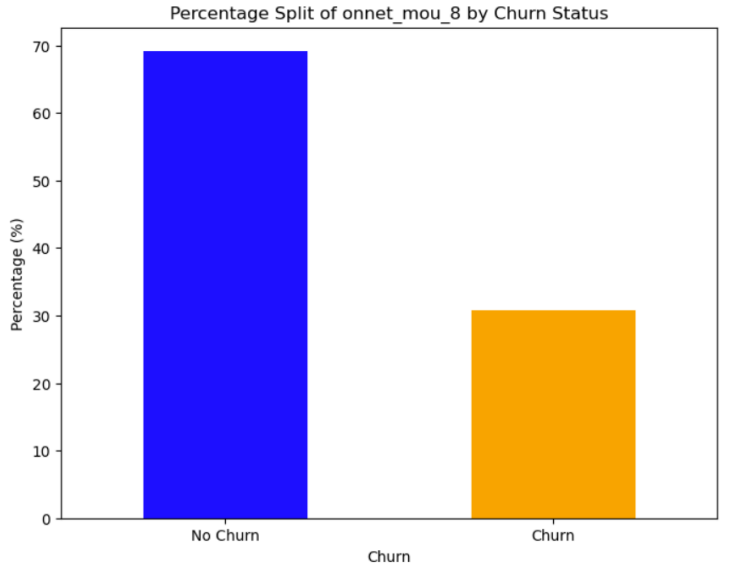


- Average Recharge Amount : Higher recharge amount customers are less likely to be churned.

Step 5: EDA - Bivariate analysis

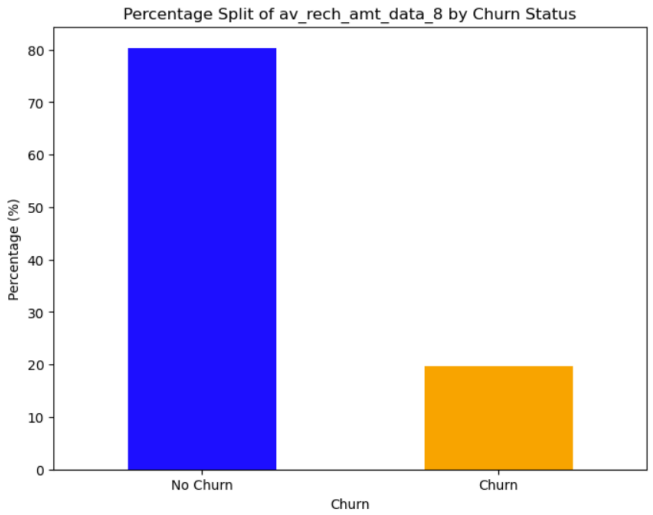


- Churn rate on the basis whether the customer decreased her/his ARPU in action month



- Churn rate on the basis whether the customer decreased her/his MOU in action month

Step 5: EDA -Bivariate analysis



- Analysis of churn rate by the average recharge amount in action month

Step 6: Test-Train Split of Dataset and Scaling

- The dataset was split into train and test sets using train_test_split.
- 80% of the data was allocated for training, and 20% for testing the model performance.
- random_state was set to 42 for having consistency across runs.
- Data Transformation using Scaling (Min Max Scaler)

Step 7: Data Modelling and Model Evaluation

- Class Imbalance is being fixed using the SMOTE Method
- Used normal Logistic Regression to check for p-values
- Used Logistic Regression using Recursive Feature Elimination (RFE) method to handle multicollinearity
- Assessing the model with StatsModels
- Creating a dataframe with the actual churn flag and the predicted probabilities
- Creating new column 'churn_pred' with 1 if Churn_Prob > 0.8 else 0
- Check for the VIF values of the feature variables
- Plotting the ROC Curve
- Finding Optimal Cutoff Point – Optimal cutoff point was found to be 0.53
- Precision and recall tradeoff
- Making predictions on the test set
- ROC Curve for the test set

Step 8: Model Selection and Comparison

Logistic regression

Train set
- Accuracy = 79%
- Sensitivity = 79%
- Specificity = 79%

Test set
- Accuracy = 80%
- Sensitivity = 77%
- Specificity = 80%

Logistic regression with PCA

Train set
- Accuracy = 91%
- Sensitivity = 09%
- Specificity = 99%

Test set
- Accuracy = 92%
- Sensitivity = 11%
- Specificity = 99%.

Decision tree with PCA

Train set
- Accuracy = 93%
- Sensitivity = 98%
- Specificity = 33%

Test set
- Accuracy = 91%
- Sensitivity = 97%
- Specificity = 21%

Step 9: Final Model Selection

We are considering the Logistic Regression as a good model to predict the Churn with 80% Accuracy

The Logistic Regression model (no PCA) has a good mix of Accuracy, Specificity and Sensitivity values when compared with other models

Step 10: Recommendation and Insights

Insights

- 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’ charges in July and ‘incoming others’ in August are less.
- Also, the customers whose increase in value-based cost in the action phase are more likely to churn. Hence, these customers may be a good target to provide offer.
- Customers with monthly 2g usage decrease 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.

Recommendations

- Telecom company needs to pay attention to the roaming rates. They need to provide good offers to the customers who are using services from a roaming zone.
- The company needs to focus on the STD and ISD rates. Perhaps, the rates are too high. Provide them with some kind of discounted STD and ISD packages.
- To address the above stated issues, it is recommended to design the product/solutions based on the customer feedback, query and complaint data.

Thank you!