Domain-Oriented Case study

Telecom Customer Churn Classification

Team:-

Ms. Anushka Saxena

Mr. Naman Arora

Ms. Archana Patil

Problem Statement

Business problem overview

- 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, we 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 and 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 we 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

- There are various ways to define churn, such as:
- **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 we 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, we will use the **usage-based** definition to define churn

High-value Churn

- In the Indian and the southeast Asian market, approximately 80% of revenue comes from the top 20% customers (called high-value customers). Thus, if we can reduce churn of the high-value customers, we will be able to reduce significant revenue leakage.
- In this project, we will define high-value customers based on a certain metric (mentioned later below) and predict churn only on high-value customers.

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 customer lifecycle :

1. **The ‘good’ phase:** In this phase, the customer is happy with the service and behaves as usual.
2. **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 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.)
3. **The ‘churn’ phase:** In this phase, the customer is said to have churned. We 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 us for prediction. Thus, after tagging churn as 1/0 based on this phase, we discard all data corresponding to this phase.

In this case, since we are working over a four-month window, the first two months are the ‘good’ phase, the third month is the ‘action’ phase, while the fourth month is the ‘churn’ phase.

Preprocessor data

Data preparation steps

- Features ‘night_pck_user’, ‘fb_user’ were converted into ‘category’
- Missing values in all variables were calculated. Missing values were imputed based on: Continuous imputed with Median and Categorical imputed with Mode
- Certain input variables that were not required for analysis were dropped as they had:
  - High missing values - ‘date_of_last_rech_data’
  - They had high low variance - ‘circle_id’
  - They were indicators - ‘mobile number’
  - Certain feature transformation tasks were also done, such as ‘aon_years’ was changed to ‘days’
- Defining and filtering of high value customers:
  - Columns related to ‘recharge amount’ were combined and to segregate high spenders by filtering customers who have spent more than 70% percentile. Remaining data came to 30,001.

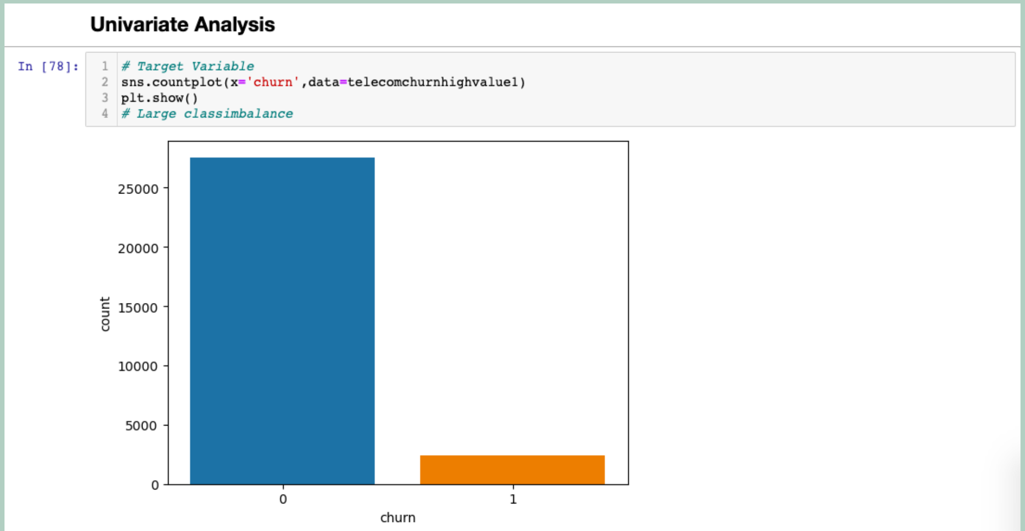
Exploratory Data Analysis

Following tasks were performed:

- Outlier identification & removal
- Univariate analysis
- Bivariate analysis
- Data visualisation

Univariate analysis

- Countplot of target variable i.e. ‘churn’ helped us identify class imbalance



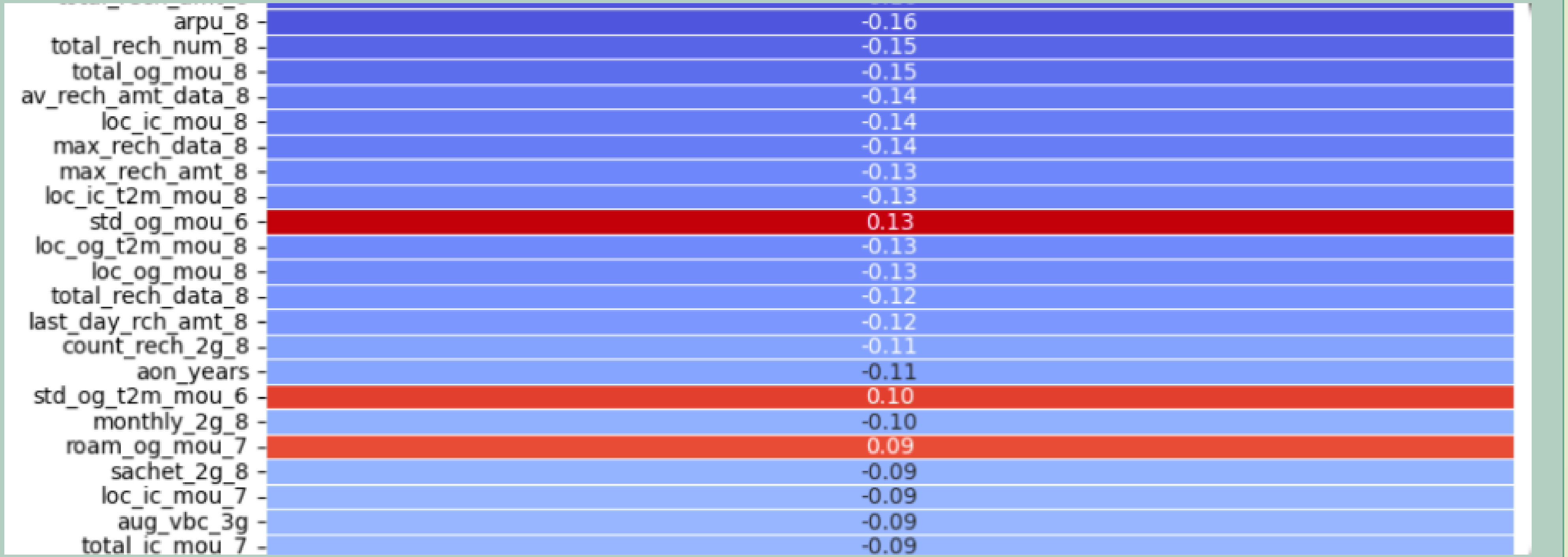
Univariate Analysis

In [78]:

```
# Target Variable
sns.countplot(x='churn',data=telecomchurnhighvalue1)
plt.show()
# Large classimbalance
```

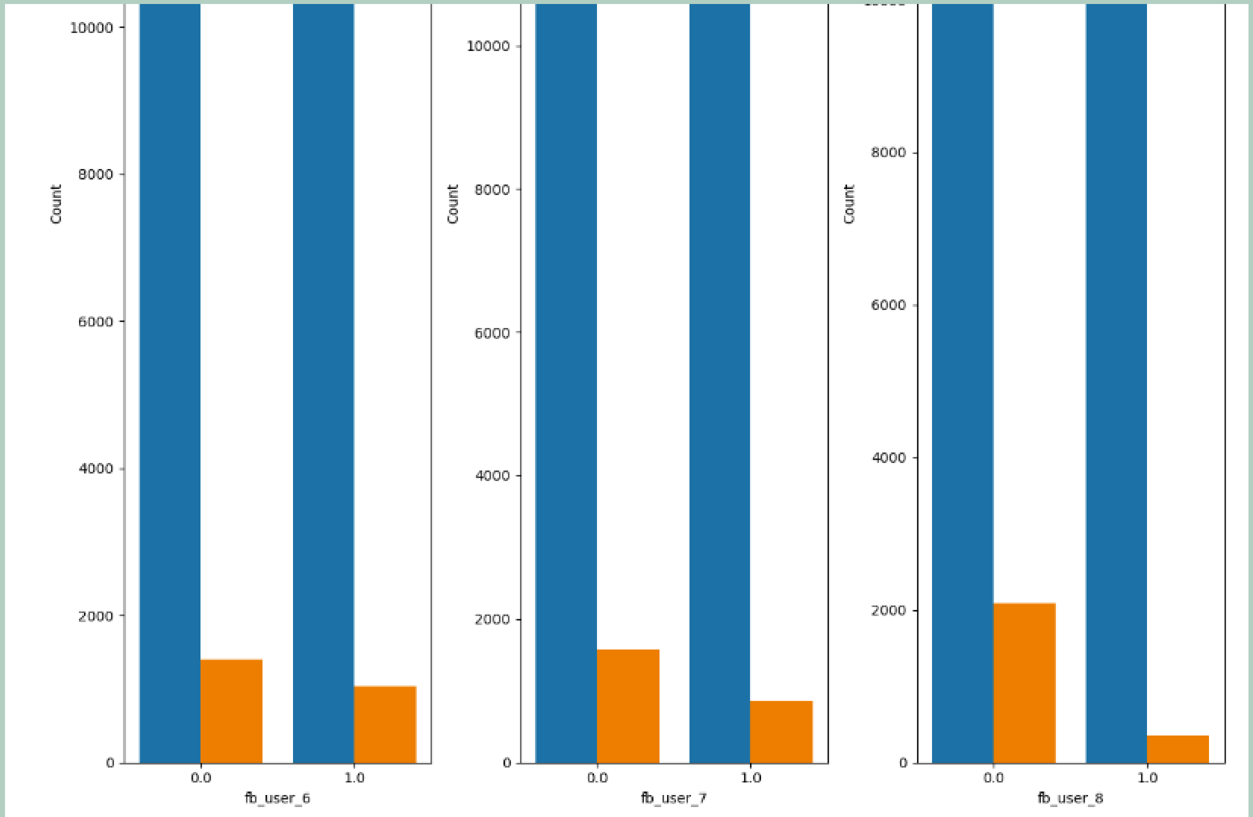
Exploratory Data Analysis

Heatmap to explore correlation between input variables and target variable



Exploratory Data Analysis

Bivariate count plot to see comparison of target variable with category input variables



Telecom Churn-Machine Learning Model

- **Feature Engineering**: Used the imbalanced-learn library to handle class imbalance in our dataset. Specifically, applied Random Under Sampling to equate the number of instances in each class, thereby ensuring that the model does not favor the majority class, didn't use random-oversampling so as to avoid the problem of overfittting
- **Imbalanced Data Handling**: Used the imbalanced-learn library to handle class imbalance in our dataset. Specifically, applied Random Under Sampling to equate the number of instances in each class, thereby ensuring that the model does not favor the majority class, didn't use random-oversampling so as to avoid the problem of overfittting
- **Normalization**: Utilized MinMaxScaler from sklearn to normalize the numerical data in the dataset. This ensures that all numerical features have a similar scale, thereby preventing any single feature from disproportionately influencing the model.
- **Model Building**: Experimented with three classification models- Logistic Regression, Decision Tree Classification and Random Forest Classification, explored the non linear models such as Decision Tree Classification and Random Forest Classification as the heatmap has brought out that there is a weak correlation between the target variable and the independent variables

Telecom Churn-Machine Learning Model-Logistic Regression

- **Recursive Feature Elimination** (RFE): Implemented RFE with Cross-Validation (RFECV) and Logistic Regression as the estimator to identify and select the most predictive features. This helps to enhance the model's performance by reducing overfitting and improving computational efficiency. rfecv.n_features_ were 102.

- **Multicollinearity Check**: Applied the Variance Inflation Factor (VIF) from statsmodels to check for multicollinearity among features. This helps to avoid overfitting and improve model interpretability, Dropped variables with high VIF and high p value in various steps

- **Model Evaluation**: Undertook model evaluation on the basis of metrics such as accuracy, precision, recall, area under curve for basic model without accounting for multicollinearity and one in which multicollinearity was accounted for both the test and train dataset

Telecom Churn-Machine Learning Model-Decision Tree Classification and Random Forest Classification

- Randomised Search CV was used for hyper parameter tuning related to number of estimators for Random forest Classifier and other parameters such as max depth, max features, min samples leaf, Randomised Search CV was used in contrast to Grid Search CV as it is computationally intensive.
- Model Evaluation: : Undertook model evaluation on the basis of metrics such as accuracy, precision, recall, area under curve for both test and train data

Model Evaluation

Model	Logistic Regression after RFECV		Logistic Regression after RFECV and addressing Multicollinearity		Decision Tree Classifier best estimator Randomised Search		Random Forest Classifier Randomised Search	
	Train	Test	Train	Test	Train	Test	Train	Test
Accuracy	80.67	78.20	78.50	77.02	85.14	81.62	87.77	88.27
Precision	79.91	24.52	78.25	23.40	86.28	27.75	90.77	39.28
Recall	81.92	80.87	78.94	80.32	83.55	78.55	84.08	81.15
F1 score	80.90	37.64	78.59	36.25	84.89	41.01	87.30	52.94

- In the context of predicting customer churn in the telecom industry, **recall is the most important metric because it represents the ability of the model to correctly identify customers at risk of churning.** Prioritizing recall ensures that we minimize the risk of missing out on any churning customers (false negatives), which is crucial given that the cost of losing a customer (lost revenue and the high cost of acquiring a new customer) is significantly higher than the cost of mistakenly trying to retain a customer who isn't at risk of churning (false positives).

Results and Recommendations

- **Logistic Regression Equation**

- churn=1.1975+(16.7762)*roam_ic_mou_7+(-5.8984)*loc_og_t2m_mou_6+(-6.4484)*loc_og_t2f_mou_6+(-3.0960)*loc_ic_mou_6+(-24.6533)*std_ic_t2t_mou_8+(-7.0222)*spl_ic_mou_8+(9.2164)*total_rech_num_7+(-28.0607)*total_rech_num_8+(30.0569)*total_rech_amt_6+(-19.8325)*last_day_rch_amt_8+(-4.1465)*count_rech_2g_8+(-1.3053)*fb_user_8+(-1.4762)*aon_years

- Logistic Regression has a high recall score in comparison to the decision tree classifier and the recall score of random forest vs logistic regression is not very different. We would be interpreting the results based on logistic regression.

Results and Recommendations

- roam_ic_mou_7: Roaming incoming minutes of usage in the 7th month has a strong positive relationship with churn. This suggests that customers who use more roaming services are more likely to churn. The telecom firm could consider offering better roaming plans or improving roaming service quality to retain these customers.
- total_rech_num_8: The total number of recharges in the 8th month has a strong negative relationship with churn. This indicates that customers who recharge less frequently in the 8th month are more likely to churn. To mitigate this, the firm could implement targeted marketing campaigns to encourage customers to recharge more frequently, or offer special deals on recharges in this period.
- total_rech_amt_6: The total recharge amount in the 6th month has a positive impact on customer retention. This suggests that customers who recharge more (higher amounts) in the 6th month are less likely to churn. Therefore, promoting higher-value recharge packages could be beneficial.
- fb_user_8: Whether the customer is a Facebook user in the 8th month has a negative correlation with churn. This suggests that customers who are not Facebook users in the 8th month are more likely to churn. It could be beneficial to explore partnerships with Facebook or offer special data packages for Facebook usage to improve customer retention.
- aon_years: The age of the customer's account (in years) negatively impacts churn, meaning customers with longer relationships with the company are less likely to churn. This emphasizes the importance of building long-term relationships with customers, possibly through loyalty programs or long-term benefits.

Thank You