Executive Summary

This data science project aims to analyze and predict user churn for a Direct-to-Home (DTH) company. By examining user demographics, service usage, billing, and customer service data, we identified key churn factors such as high service costs and poor customer interactions. Using machine learning models, particularly Random Forest, we achieved predictive accuracy with an ROC-AUC score of **“0.91”**. Recommendations include flexible pricing with attractive discounts, improved service reliability, and proactive customer engagement. Implementing these strategies can reduce churn, enhance customer satisfaction, and drive sustainable growth. Future work involves real-time prediction, advanced techniques, and continuous model updates.

Table of Contents

Customer Churn Prediction for DTH industry

1. Introduction

1.1 Project Background/Motivation

Customer churn, the phenomenon where users discontinue a service, presents a significant challenge for businesses across various industries. For Direct-to-Home (DTH) companies, churn not only impacts revenue but also affects long-term growth and market competitiveness. Understanding the underlying reasons behind customer churn and predicting it accurately can empower DTH companies to devise effective retention strategies, thereby ensuring a steady revenue stream and enhanced customer loyalty.

As industries across both products and services transition to subscription models, understanding and managing customer churn has become critical. Accurately predicting churn through data analysis allows businesses to implement proactive strategies to retain customers and sustain revenue streams. By leveraging data insights, companies can identify at-risk customers and take targeted actions to address their concerns before they decide to leave. This proactive approach not only helps in preventing churn but also ensures a steady customer base, ultimately supporting long-term business stability and growth in a competitive subscription-based market.

1.2 Understanding DTH Companies

Direct-to-Home (DTH) is a satellite television broadcasting service where signals are transmitted directly to a subscriber’s premises. Unlike traditional cable TV services that require a local distribution network, DTH provides digital satellite television signals directly to the user’s dish antenna installed at their location. This method ensures a higher quality of signal reception and a broader range of channels and services. In India, the DTH market has grown significantly, driven by the country's increasing digitalization, a diverse range of content offerings, and competitive pricing strategies.

1.3 Market Value and Customer Base in India

The DTH market in India is substantial, with millions of subscribers across the country. As of recent reports, the Indian DTH market was valued at approximately USD 2.9 billion in 2023 and is expected to grow at a compound annual growth rate (CAGR) of 9.8% from 2023 to 2028. The customer base is diverse, including urban and rural areas, catering to various preferences with numerous regional and international channels. Major players in the Indian DTH market include Tata Sky, Dish TV, Airtel Digital TV, and Sun Direct, which collectively serve over 70 million active subscribers.

Despite its significant market presence, the DTH industry faces stiff competition from emerging Over-the-Top (OTT) platforms like Netflix, Amazon Prime Video, and Disney+ Hotstar. OTT services provide several advantages over traditional DTH, including On-demand content, personalized recommendation based on algorithms etc.

These advantages make OTT services increasingly popular, leading to higher churn rates in the DTH sector as customers shift towards these modern, convenient alternatives.

2. Problem Statement

In the competitive landscape of DTH services, retaining customers is more cost-effective than acquiring new ones. However, predicting which customers are likely to churn is a complex task, influenced by numerous factors such as service quality, pricing, customer service interactions, and user behavior patterns. The goal of this project is to identify the major churn factors and develop a robust predictive model that can identify customers at high risk of churning, allowing the company to proactively address the issues and improve retention rates.

3. Data Description

Raw data provided has a total of 19 variables in which 9 are categorical in nature and 10 continuous. The total number of observations is 11260.

#	Column	non-null Count	Dtype
1	AccountID	10982 non-null	int64
2	Churn	10982 non-null	int64
3	Tenure	10766 non-null	float64
4	City_Tier	10870 non-null	float64
5	CC_Contacted_LY	10881 non-null	float64
6	Payment	10873 non-null	object
7	Gender	10874 non-null	object
8	Service_Score	10884 non-null	float64
9	Account_user_count	10540 non-null	float64
10	account_segment	10885 non-null	object
11	CC_Agent_Score	10866 non-null	float64
12	Marital_Status	10770 non-null	object
13	rev_per_month	10193 non-null	float64
14	Complain_ly	10625 non-null	float64
15	rev_growth_yoy	10979 non-null	float64
16	coupon_used_for_payment	10979 non-null	float64
17	Day_Since_CC_connect	10625 non-null	float64
18	cashback	10512 non-null	object
19	Login_device	10762 non-null	object

From the initial investigations of the data it is clear that there are certain discrepancies in the data from incompleteness to inaccuracy which suggest the requirement of robust data wrangling to be conducted.

Assumptions :

1. Other than Tenure (99 months) and Coupons used for payment (> 12), we have not removed any outliers.
2. All the subscriptions are monthly.

4. Methodology Overview

1. **Data Collection and Preprocessing:** Gather comprehensive data on customer demographics, service usage, billing history, and customer service interactions. Clean and preprocess this data to ensure it is suitable for analysis.
2. **Exploratory Data Analysis (EDA):** Perform EDA to uncover initial insights, understand data distributions, and identify potential patterns and correlations that may contribute to customer churn.
3. **Model Development:** Develop and compare multiple machine learning models, including Logistic Regression, Decision Trees, Random Forest to predict customer churn. Evaluate these models based on accuracy, precision, recall, and ROC-AUC score.
4. **Model Validation and Testing:** Validate the chosen model using cross-validation techniques to ensure its robustness and generalizability. Test the model on a separate dataset to assess its performance in real-world scenarios.
5. **Actionable Insights and Recommendations:** Based on the model's predictions and the insights derived from the analysis, provide strategic recommendations to the DTH company. These may include improving service quality, optimizing pricing strategies, and enhancing customer engagement efforts.

5. Methodology

5.1 Data Collection and Preprocessing

Handling wrong data type:

- All variable data types were verified and validated, in which "*cashback*" data type was wrongly defined as an object/string variable, thus the same was converted into a float.

Handling Null values: (Refer ***Exhibit 2- Table 2***).

- Total number of 3812 "Null" values were observed in the dataset.
- Case to case columns were evaluated and based on the context of information rows with nulls where either deleted or column respective values were replaced with mean or mode.
- Example *'Payment', 'Gender'* was concluded to be critical data without which model building will be adversely affected and replacing them with mode or median was not found to be a rational approach thus rows without this data were deleted from the dataset.

- On the other hand for *'City_Tier'* it was concluded that it makes sense to replace nulls with the majority *'City_Tier'* value (mode) as the probability of the missing city tier to be the mode value to be high.
- Similarly for missing *'cashback'* values, mean was used to replace the missing values.
- Inputting “Null” values:
  - Removing the rows pertaining to: *Tenure', 'Payment', 'Gender', 'Account_user_count', 'account_segment', 'Marital_Status', 'Login_device'.*
  - Inputting with mean: *'CC_Contacted_LY', 'Service_Score', 'CC_Agent_Score', 'rev_growth_yoy','Day_Since_CC_connect','rev_per_month','cashback','coupon_used_for_payment'.*
  - Inputting with mode : *'City_Tier','Complain_ly'.*

Handling incorrect data: (Refer ***Exhibit No 2 Table 1***)

- Beyond null values entered data was also verified to validate correctness.
- *‘Tenure’* for example has UOM as months and certain rows had a value of 99, this was judged to be incorrect data as practically there are neither DTH plans in the market nor do customers get into 99 months contracts with service providers.
- This must be a default value assignment at some data storage level, thus such rows were deleted as it will negatively influence the model.
- The dataset has various special characters like #, &, +, *, @,$ where in the rows containing such characters were removed.

Final Processed data :

Of the 19 features after data cleanup we had a total of 9578 observations compared to the original set of 11260.

5.2 Exploratory Data Analysis (EDA)/ Descriptive analysis

- Multiple descriptive analytical tools were used to find patterns and correlations from the variables.
- Univariate and Multivariate analysis was performed using Box & Whisker plots, Bar charts etc to find patterns.
- Correlation matrix study was prepared for numerical variables.
- Refer ***Exhibit No 3*** for detailed plots.

5.3 Model Development:

Preprocessing and fitting:

a. Encoding Categorical variables

In this analysis, we are converting categorical variables into factors using the ***One Hot Encoding*** method instead of labeling method. The reason for this choice is that labeling categorical variables can unintentionally imply a ranking or order among the categories, which is often not appropriate. One Hot Encoding, on the other hand, avoids this issue by creating a new binary variable for each category. This approach ensures that each category is treated independently, preserving the categorical nature of the data without imposing any sort of ordinal relationship.

Categorical variables: ['City_Tier', 'Payment', 'Gender','account_segment','Marital_Status', 'Complain_ly', 'Login_device'].

b. Splitting the data to train and test set

We are considering an ***80:20*** split in train and test data. Also, the data is bifurcated into features and target variable wherein:

·Features are: 'Tenure', 'City_Tier', 'CC_Contacted_LY', 'Service_Score', 'Account_user_count', 'CC_Agent_Score','rev_per_month','Complain_ly','rev_growth_yoy','coupon_used_for_payment', 'Day_Since_CC_connect','cashback','Payment_Credit Card','Payment_Debit Card', 'Payment_E wallet','Payment_UPI', 'Gender_Male', 'account_segment_Regular', 'account_segment_Regular Plus', 'account_segment_Super', 'account_segment_Super Plus', 'Marital_Status_Married', 'Marital_Status_Single', 'Login_device_Mobile'.

(Account ID has been removed since it is insignificant for modeling).
Target variable is Churn status.

c. Scaling the data

The data is standardized such that each feature has a mean of 0 and a standard deviation of 1. Same scaling has been applied to the test data using the mean and standard deviation calculated from the training data.

d. Fitting the model

The following models have been considered for churn classification:

- Logistic regression
- Random forest
- Decision tree
- Pruning the decision tree:
    - Decision tree was pruned considering 3 parameters : max_depth': 20,'min_samples_leaf': 1, 'min_samples_split': 2.
    - The cross validation accuracy was found to be 93%.

5.4 Model Cross validation

This being a classification problem the cross validation of models will be done based on their Precision, Recall and F1 scores.

6. Results & Analysis

6.1 Univariate

Categorical variable analysis

- **Tenure distribution:**
    - 75 % of users have opted for tenure which is in and around the 1 year mark or lower.
    - Thus, the possibility of yearly churn is high as within a year 75 % of users will re-think the decision to continue subscription or not.
    - Thus, it's very critical to predict churn & take necessary action to prevent churn.
    - Tenure had 4 outliers.
- **Customer care contacted in last year:**
    - 50% of customers contacted customer care less than 10 times in a year, that's less than a call a month, which is a decent level of quality of service.
    - 50% of customers contacted customer care less than 10 times in a year, that's less than a call a month, which is a decent level of quality of service.
    - There are 25% of users who have contacted the customer care 10 to 15 times a year and quite a few customers called 40 to 120 times a year which is quite alarming.
    - This group including the outliers is a high-risk group that can churn.
- **Service score:**
    - 25% of customers have given a score of less than 2 which can lead to churn.
    - Median and third quartile coincide with each other.
- **Account_user account:**
    - There are a minimum of 2 users who use an account minus the few lower outliers.
    - Median and third quartile coincide with each other.
- **CC Agent Score:**
    - 25% of customers have given a score of less than 2.
    - 50% of customers have given a score of less than 3.
    - 50% of customers have given a score of higher than 3.
- **Coupon used:**
    - 100 % of customers have used 3 or less coupons.
    - Customers who have utilized coupons are in the outliers.
    - It seems coupons are not reaching the majority of customers or they are too difficult to redeem.
- **Day_Since_CC_connect:**
    - 75% of contacts lie below 9 days.
- **Cashback:**
    - 75% customers have got a minimum cashback of below 200 and 25 % of customers have got a cashback of below 150.
    - There is a high number of high end outliers when it comes to cashback.
- **Payment:**

- The maximum churn happens on those customers who pay through debit and credit card. Organization needs to analyze why this happens and some strategies to be in place like cashback's / discounts on next recharge, etc.

6.2 Multivariate (Refer ***Exhibit No 3***)

Gender,City tier & Payment mode there seems to have the same proportion of churn.

Segment wise Churn:

- Regular plus category although has lesser accounts than Super,it has churned more than Super category(Highest subscription).

Marital status:

- Single people seem to churn more than married or divorced while comparing proportions.

Complain_ly:

- 30% of customers have raised complaints last year.
- Out of which more than 50% of customers had churned.
- Whereas customers who haven't complained have only churned around 15%.

Correlation results of numerical features:

- Rarely any negative correlation was observed.
- Highest positive correlation was found between *Days_since_cc_connect* and *Coupon_used for_payment*(0.35).
- Followed by *Days_since_cc_connect* and *cashback*(0.33).
- Account user-count and Service_core have high positive correlation(0.32).
- Churn has a negative correlation with "*Day_since_CC_connect*" and "*cashback*"(-0.14).

6.3 Model Analysis

Factors Positively Related to Churn (Decrease Likelihood of Churn)

Feature	Coefficient	Odds Ratio	Inference	Business Insights
Complain_ly_1.0	0.801476	2.228828	Customers who complained last year are significantly more likely to churn.	Take immediate action to address complaints, offer coupons, and provide cashbacks to retain these customers.
rev_per_month	0.53914	1.714531	Higher monthly revenue is associated with higher odds of customer churn.	Review pricing and compare with competitors, as customers may churn if they find similar services at a lower price elsewhere.
account_segment_Regular	0.416936	1.517305	Regular segment customers are more likely to churn.	Create more attractive plans for regular customers or promote the benefits of upgrading to Plus or Super plans.
Account_user_count	0.39298	1.481389	Accounts with more users are more likely to churn.	Consider limiting the number of users per account, as it may be challenging to satisfy different individuals.

Marital_Status_Single	0.354521	1.425498	Single customers are more likely to churn.	No significant insight.
CC_Agent_Score	0.346193	1.413675	Higher customer care agent scores increase the likelihood of customer churn.	Consider this an anomaly.
City_Tier_3.0	0.314572	1.369673	Tier 3 customers have a higher likelihood of churning.	Investigate reasons for dissatisfaction in Tier 3 cities and take corrective measures, while maintaining consistent service quality in Tier 1.
CC_Contacted_LY	0.299375	1.349016	Customers who contacted customer care last year are more likely to churn.	Improve service quality.
coupon_used_for_payment	0.286529	1.331796	Customers using coupons for payment are slightly more likely to churn.	Consider this an anomaly.
City_Tier_2.0	0.164211	1.178463	Tier 2 customers have a higher likelihood of churning.	Investigate reasons for dissatisfaction in Tier 2 cities and take corrective measures, while maintaining consistent service quality in Tier 1.
Gender_Male	0.188769	1.207762	Male customers are more likely to churn.	No significant insight.
account_segment_Super Plus	0.029572	1.030014	Super Plus segment customers have a negligible effect on churn likelihood.	No significant insight.

Factors Negatively Related to Churn (Decrease Likelihood of Churn)

Feature	Coefficient	Odds Ratio	Inference	Business Insights
account_segment_Super	-0.763381	0.466088	Super segment customers are much less likely to churn.	Promote Regular Plus or Super plans to new customers to reduce churn risk.
Tenure	-1.939636	0.143756	Longer tenure significantly decreases the odds of a customer churning.	Introducing attractive discounts for long-term plans can reduce the possibility of churn.
cashback	-0.369735	0.690917	Offering cashback reduces the odds of customer churn.	Provide cashbacks to high-risk customers to reduce churn.
account_segment_Regular Plus	-0.23675	0.789189	Regular Plus segment customers are less likely to churn.	Encourage customers to opt for Regular Plus or Super plans as they have lower churn rates.
Payment_Credit Card	-0.28539	0.751721	Using credit cards for payment slightly reduces the odds of customer churn.	No significant insight.
Payment_Debit Card	-0.216293	0.805499	Using debit cards for payment slightly reduces the odds of customer churn.	No significant insight.
Login_device_Mobile	-0.210746	0.80998	Customers using mobile devices to log in are less likely to churn.	No significant insight.

Day_Since_CC_connect	-0.216646	0.805215	More days since the last customer care connection reduces the odds of customer churn.	Fewer complaints indicate less churn risk.
Payment_UPI	-0.18036	0.83497	Using UPI for payment reduces the odds of customer churn.	No significant insight.
Service_Score	-0.07768	0.925261	A higher service score slightly reduces the odds of customer churn.	Improve service quality.
Marital_Status_Married	-0.097166	0.907405	Married customers have a slightly reduced likelihood of churning.	No significant insight.
rev_growth_yoy	-0.123319	0.883982	Higher year-over-year revenue growth slightly reduces the odds of customer churn.	Satisfied customers are willing to spend more on service, leading to reduced churn and increased revenue YOY.

6.3.1 Logistic Regression Model (Training data)

Metric	Positive Class (Churn)	Negative Class (Non-Churn)	Overall	Interpretation
Accuracy	-	-	0.9	The model correctly predicts the outcome 90% of the time, indicating good overall performance.
Precision	0.77	0.91	0.89	Precision for churn (77%) means 77% of predicted churns are actual churns. For non-churn (91%), it means 91% of predicted non-churns are actual non-churns.
Recall	0.54	0.97	0.75	Recall for churn (54%) indicates the model identifies 54% of actual churned customers. For non-churn (97%), it means the model identifies 97% of actual non-churned customers.
F1-Score	0.63	0.94	0.79	F1-Score for churn (0.63) balances precision and recall, showing moderate performance. For non-churn (0.94), it reflects a strong balance of precision and recall.
ROC AUC Score	-	-	0.75	AUC score of 0.75 indicates the model's good ability to distinguish between churned and non-churned customers.

Refer ***Exhibit No: 5***

6.3.2 Random Forest Model

Metric	Positive Class (Churn)	Negative Class (Non-Churn)	Overall	Interpretation
Accuracy	-	-	1	The model correctly predicts the outcome 100% of the time, indicating excellent performance.
Precision	1	1	1	Precision for churn (100%) means all predicted churns are actual churns. For non-churn (100%), all predicted non-churns are actual non-churns.
Recall	1	1	1	Recall for churn (100%) indicates the model identifies all actual churned customers. For non-churn (100%), it means the model identifies all actual non-churned customers.
F1-Score	1	1	1	F1-Score for churn (1.00) reflects a perfect balance between precision and recall. For non-churn (1.00), it also shows a perfect balance.
ROC AUC Score	-	-	1	A ROC AUC score of 1.00 indicates the model has a perfect ability to distinguish between churned and non-churned customers.

Refer ***Exhibit No: 5***

6.3.3 Decision Tree model

Metric	Positive Class (Churn)	Negative Class (Non-Churn)	Overall	Interpretation
Accuracy	-	-	1	The model correctly predicts the outcome 100% of the time, indicating excellent performance.
Precision	1	1	1	Precision for churn (100%) means all predicted churns are actual churns. For non-churn (100%), all predicted non-churns are actual non-churns.
Recall	1	1	1	Recall for churn (100%) indicates the model identifies all actual churned customers. For non-churn (100%), it means the model identifies all actual non-churned customers.
F1-Score	1	1	1	F1-Score for churn (1.00) reflects a perfect balance between precision and recall. For non-churn (1.00), it also shows a perfect balance.
ROC AUC Score	-	-	1	A ROC AUC score of 1.00 indicates the model has a perfect ability to distinguish between churned and non-churned customers.

Refer ***Exhibit No: 5 Figure 30***

6.3.4 Pruned Decision Tree model

Metric	Positive Class (Churn)	Negative Class (Non-Churn)	Overall	Interpretation
Accuracy	-	-	1	The model correctly predicts the outcome 100% of the time, indicating excellent performance.
Precision	1	1	1	Precision for churn (100%) means all predicted churns are actual churns. For non-churn (100%), all predicted non-churns are actual non-churns.
Recall	0.99	1	1	Recall for churn (99%) indicates the model identifies 99% of actual churned customers. For non-churn (100%), it means the model identifies all actual non-churned customers.
F1-Score	1	1	1	F1-Score for churn (1.00) reflects a perfect balance between precision and recall. For non-churn (1.00), it also shows a perfect balance.
ROC AUC Score	-	-	1	A ROC AUC score of 1.00 indicates the model has a perfect ability to distinguish between churned and non-churned customers.

Refer ***Exhibit No : 5 Figure 32***

6.4 Model performance on Positive class (Churn) for training data

Model	Precision (Churn)	Recall (Churn)	F1-Score (Churn)	ROC AUC Score
Logistic Regression	0.77	0.54	0.63	0.75
Random Forest	1	1	1	1
Decision Tree	1	1	1	1
Pruned Decision Tree	1	0.99	1	1

6.5 Model performance comparison test vs training data

Model	Dataset	Precision (Churn)	Recall (Churn)	F1-Score (Churn)	ROC AUC Score
Logistic Regression	Train	0.77	0.54	0.63	0.75
	Test	0.78	0.53	0.63	0.75
Random Forest	Train	1	1	1	1
	Test	0.98	0.89	0.93	0.94
Decision Tree	Train	1	1	1	1
	Test	0.89	0.91	0.9	0.94
Pruned Decision Tree	Train	0.99	0.96	0.98	0.98
	Test	0.9	0.87	0.88	0.93

Both the Decision Tree and Random Forest show strong performance on the training data, but Random Forest has a slight edge in generalization due to its ensemble nature. This approach helps reduce overfitting by averaging the predictions of multiple trees, providing more reliable and consistent results on the test data. While both models perform well, ***Random Forest offers better precision, making it the recommended choice for predicting customer churn.***

6.6 Model performance on test data (PCA vs Non PCA)

Model	Data	Test Precision (Churn)	Test Recall (Churn)	Test F1-Score (Churn)	Test ROC AUC Score
Logistic Regression	With PCA	0.76	0.5	0.6	0.73
	Without PCA	0.78	0.53	0.63	0.75
Random Forest	With PCA	0.98	0.68	0.8	0.84
	Without PCA	0.98	0.89	0.93	0.94
Decision Tree	With PCA	0.75	0.72	0.73	0.83
	Without PCA	0.89	0.91	0.9	0.94
Pruned Decision Tree	With PCA	0.74	0.69	0.71	0.82
	Without PCA	0.9	0.87	0.88	0.93

PCA has degraded model performance due to the loss of important information when reducing dimensionality. Another reason could be Oversimplification wherein only few components are retained, potentially eliminating features critical for distinguishing the positive class.
Thus we have considered refitting the model based on Feature importance
(Refer ***Exhibit No 5 Figure no: 33***)

6.7 Feature Influence:

- Tenure remains a critical feature across both models, indicating its strong impact on predicting churn.
- Features like *Day_Since_CC_connect, cashback, and CC_Agent_Score* show moderate importance in both models.
- Model refitting was done using only the features which have **importance score more than 0.03**

'Tenure', 'CC_Contacted_LY', 'Account_user_count', ‘CC_Agent_Score', 'rev_per_month', 'Complain_ly', 'rev_growth_yoy', 'Day_Since_CC_connect', 'cashback', 'Marital_Status_Single', 'Login_device_Mobile'.

Model	Dataset	Test Precision (Churn)	Test Recall (Churn)	Test F1-Score (Churn)	Test ROC AUC Score

Logistic Regression	**All Features**	0.78	0.53	0.63	0.75
	Important Features	0.82	0.49	0.62	0.74
Random Forest	**All Features**	0.98	0.89	0.93	0.94
	Important Features	0.96	0.83	0.89	0.91
Decision Tree	**All Features**	0.89	0.91	0.9	0.94
	Important Features	0.84	0.80	0.82	0.88
Pruned Decision Tree	**All Features**	0.9	0.87	0.88	0.93
	Important Features	0.88	0.78	0.83	0.88

The Random Forest model exhibits the most stable performance when reducing features, showing the smallest drop in key metrics. Logistic Regression shows an improvement in precision but a decline in recall and ROC AUC Score. Both Decision Trees and Pruned Decision Trees experience slight reductions in performance metrics, but they still maintain competitive scores. So, ***Random Forest appears more robust for prediction compared to other models.***

7. Interpretations & Discussions

- From the result analysis it's clear that the Random Forest model is the best option for predictions on churn.
- For predicting future churn the the accounts along with its features will be fed into the random forest model to predict churn.
- From the feature importance output we also know the features which have most impact to least impact for the model.
- Thus the company should take more care in collecting clean data with respect to the important features so that no data is lost.
- We would suggest the firm concentrate on the top 10 features from the importance output so that good quality data is acquired for future predictions.
- Based on total revenue, the top 2 segments are Super and Regular Plus, and 15% churn was for Super and 30% for regular plus.
- Thus more attention is to be provided to prevent churn in these 2 segments.
- Based on the churn predictions the firm needs to take corrective action to prevent the churn from happening.
- For this we will look into the logistic regression model outcomes and assess the feature level coefficients and odds ratio to identify the features on which corrective action is to be taken.
- All features that fall in the overlap of the high importance model of Random forest and high coefficient and odds ratio of logistic regression will be considered as most important features on which action is to be taken to prevent possible churn in the future.

8. Conclusion

- DTH is a highly oligopolistic market where all the competitors are facing stiff competition. It has a huge growth potential but due to multiple competitors customers can easily churn from one provider to another.
- The real challenge during the modeling process is to avoid model overfitting and underfitting. Both these cases will result in reduced accuracy of the classifier.
- It is observed that the *Random Forest model* is the best model as the F1 Score is higher compared to other models with respect to train and test dataset.
- For regular customers, create attractive plans or promote upgrading to Plus or Super plans.
- Promote Plus or Super plans to new customers to lower churn risk.
- Limit the number of users per account based on the plan to better meet individual needs.
- Improve service quality for those who contacted customer service last year.
- Investigate dissatisfaction in Tier 2 and Tier 3 cities, while maintaining quality in Tier 1.
- Address complaints immediately with actions like offering coupons and cashbacks.
- Review pricing and competitors to prevent customers from churning for cheaper alternatives.
- Continuously improve service quality score and maintain it.
- Offering discounts for long-term plans to reduce churn risk.
- Identify High risk customers and provide personalized offers to reduce the churn.

9. References

1. Capstone Project - Churn Prediction (kaggle.com)