CUSTOMER CHURN PREDICTION

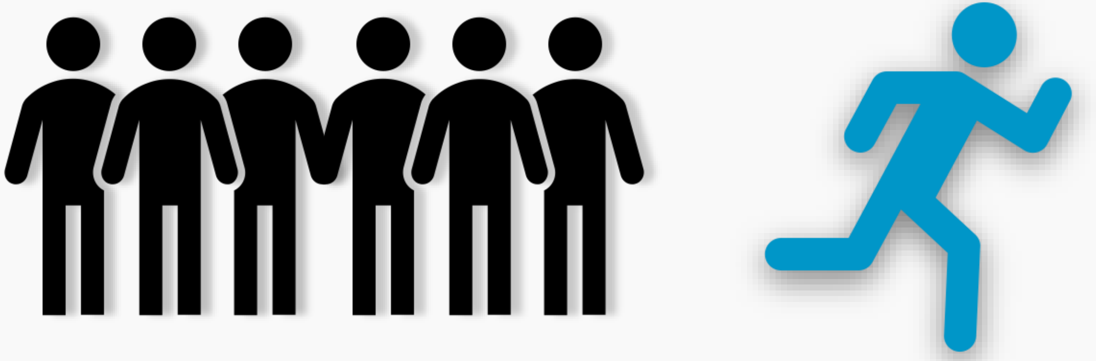
greatlearning

Mentor
Mr. Jatinder Bedi

Team Members

Harshit Agrawal
Hrishikesh Gupta
Shubham Sharma
Sanket Wakalkar
Vamsi Krishna

WHAT IS CUSTOMER CHURN ?



PROBLEM STATEMENT

greatlearning

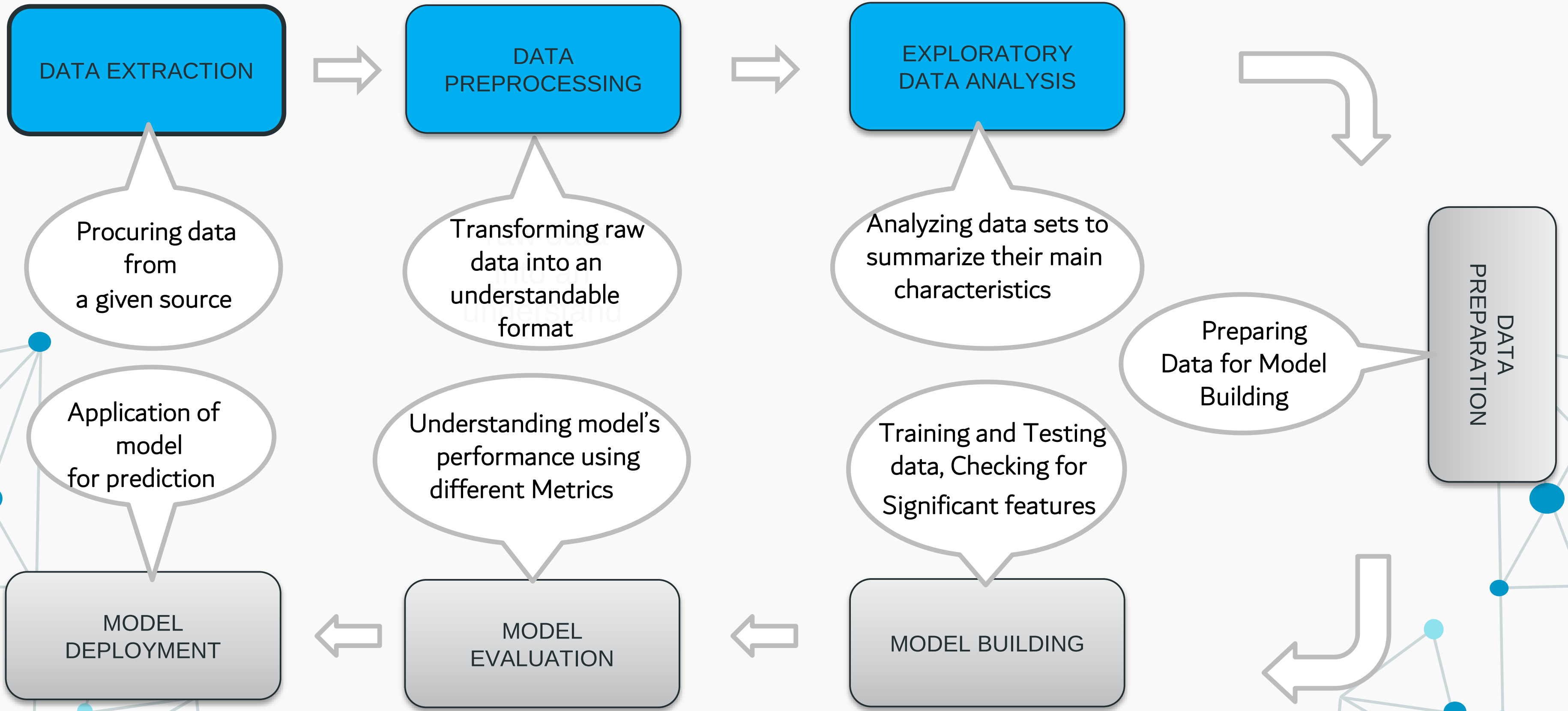
- The telecommunications industry experiences an average of 1.9% monthly and 22% annual churn rate. (SOURCE: GOOGLE)

- We know 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.

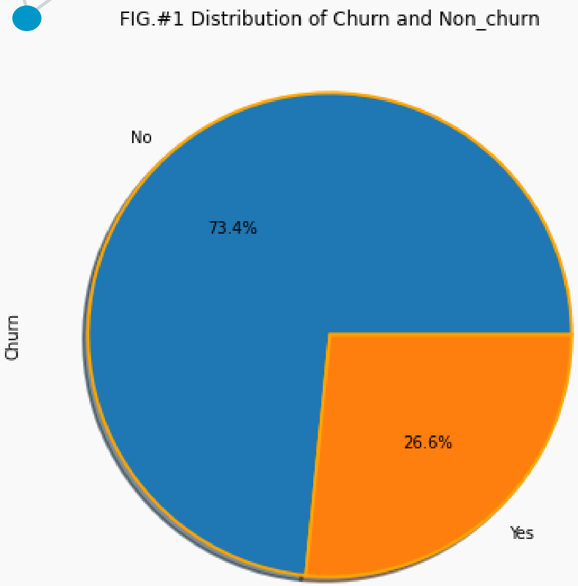
- To reduce customer churn, telecom companies need to predict which customers are at high risk of churn. In this project, we will Analyze customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn.

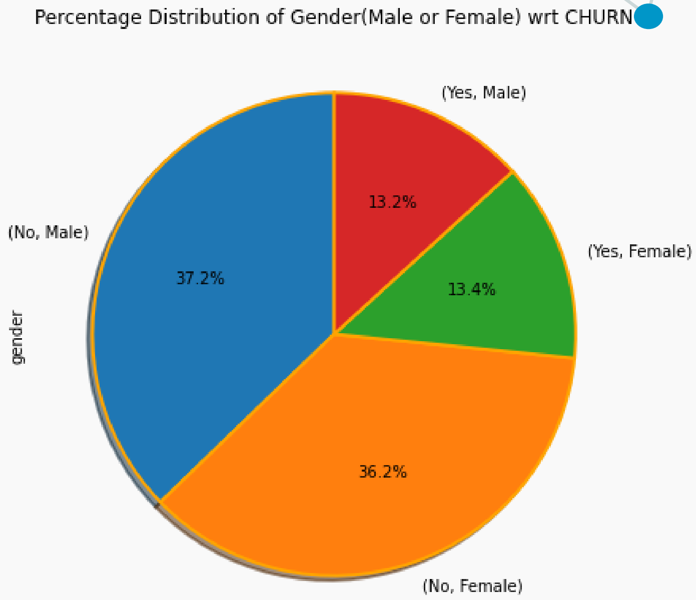
- Here our goal is to build a machine learning model that is able to predict churn of customers based on the features provided for their usage.

FLOWCHART


DATA EXTRACTION
DATA PREPROCESSING
EXPLORATORY DATA ANALYSIS
DATA PREPARATION
MODEL BUILDING
MODEL EVALUATION
MODEL DEPLOYMENT
Procuring data from a given source
Transforming raw data into an understandable format
Analyzing data sets to summarize their main characteristics
Preparing Data for Model Building
Training and Testing data, Checking for Significant features
Understanding model's performance using different Metrics
Application of model for prediction

EXPLORATORY DATA ANALYSIS



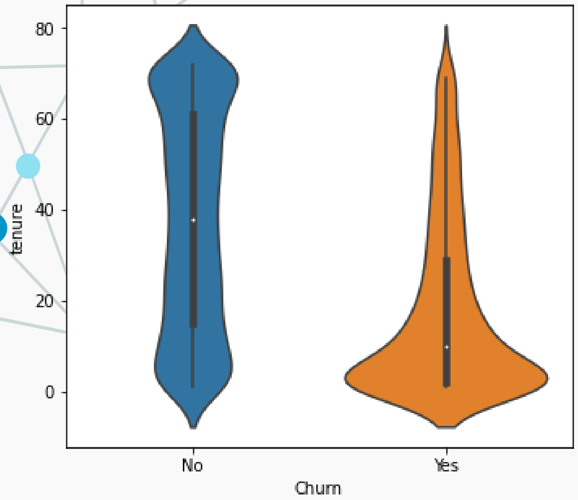


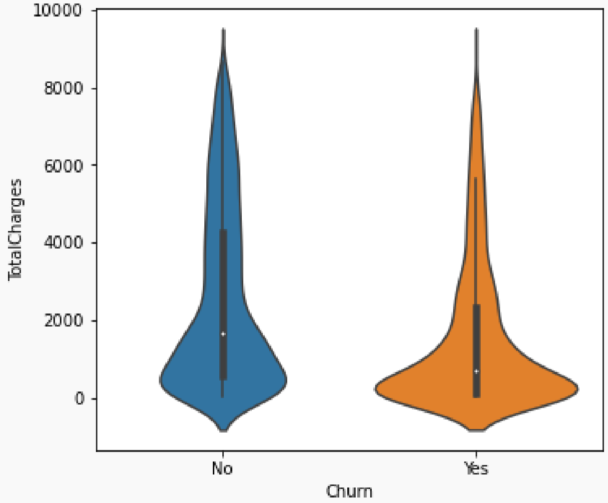
Fig#1: Pie chart here depicts the distribution of the target variable CHURN. Ratio of NO:YES is nearly 73:27

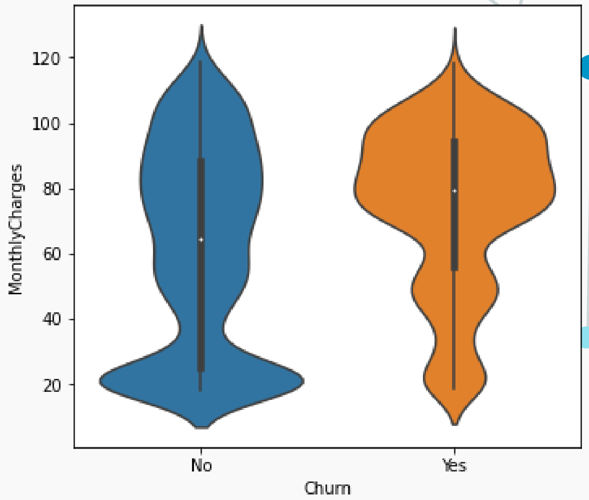
Fig#2: Depiction of gender distribution w.r.t CHURN. Gender w.r.t CHURN is equally distributed.

EXPLORATORY DATA ANALYSIS

greatlearning







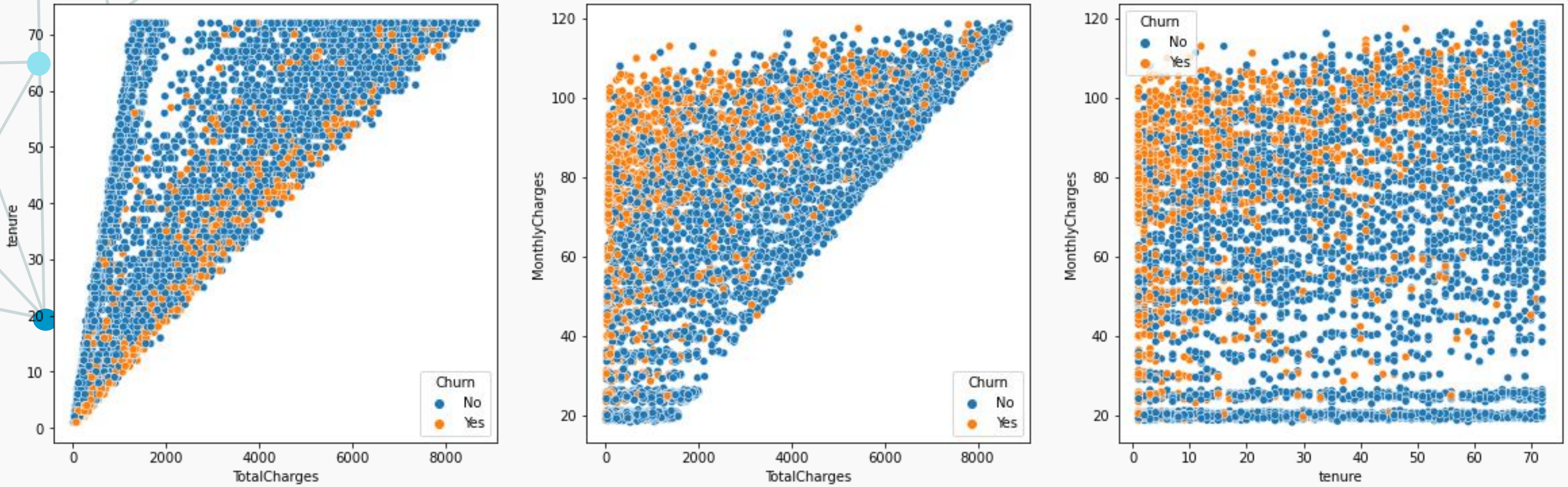
Longer the association, more loyal the customer

As total charge increases, customer becomes more associated with the company

As monthly charge increases , customer terminate the services

EXPLORATORY DATA ANALYSIS

greatlearning



- The lower the total charges and tenure, the higher the churn.
- Churn is higher for higher bands of monthly charges.
- If tenure is less, and monthly charges are more , customer is more likely to churn

EXPLORATORY DATA ANALYSIS

greatlearning

There were 11 null values in TotalCharges variable

The dataset does not have any outliers

HEATMAP shows a very high relation between tenure and total charges

Independent variables 17 categorical and 3 continuous

The data is imbalanced with ratio of NO:YES as 73:27

Gender distribution is equal for male and female w.r.t churn

Most of the customers are young people with 84% as contribution

Some columns had extra class which were not relevant

PREPROCESSING

Dummy Encoding of the categorical variables (N-1 encoding)

Scaling of Numerical variables

Label Encoding of the Target variable

Removed rows with the missing values

Dropped irrelevant column like CustomerID and also dropped Gender as it is non significant variable

Splitting of the train and test set for the further analysis

Balancing of the data using AdaSyn

LOGISTIC REGRESSION

1. Logistic regression for both balanced and imbalanced data set is performed.
2. The idea behind applying logistic regression here is to get an insight of how the data set is behaving towards the classification algorithm.
3. Data being balanced or imbalanced, we can surely say here that accuracy of not more than 0.79 is achieved.
4. Also in imbalanced data accuracy is not reliable metric so we can look to the recall score. Maximum of 0.80 is the recall score that is achieved.
5. For balanced data accuracy achieved is near about 0.74

	Model	Data Balanced/Imbalanced	Probability Cutoff	AUC Score	Precision Score	Recall Score	Accuracy Score	Kappa Score	f1-score
0	Logreg_full	Imbalanced	0.500	0.704147	0.642058	0.511586	0.794313	0.436588	0.569444
1	Logreg_full	Imbalanced	0.258	0.766789	0.519630	0.802139	0.750237	0.454734	0.630694
2	logreg_rfe	Imbalanced	0.500	0.706575	0.646532	0.515152	0.796209	0.441780	0.573413
3	logreg_rfe	Imbalanced	0.260	0.763070	0.514943	0.798574	0.746445	0.447528	0.626136
4	logreg_bal	Balanced	0.500	0.736549	0.505535	0.732620	0.738389	0.413816	0.598253
5	logreg_bal	Balanced	0.450	0.741569	0.487208	0.780749	0.723223	0.405267	0.600000
6	logreg_bal_forward	Balanced	0.500	0.735658	0.504926	0.730838	0.737915	0.412462	0.597232
7	logreg_bal_forward	Balanced	0.470	0.743968	0.497123	0.770053	0.731754	0.415233	0.604196

greatlearning

DECISION TREE

Further the data is subjected to Decision tree modelling

	Model	Data Balanced/Imbalanced	Precision Score	Recall Score	Accuracy Score	Kappa Score	f1-score
0	Decision_tree(Gini)	Imbalanced	0.462385	0.449198	0.714692	0.262431	0.455696
1	Decision_tree(Entropy)	Imbalanced	0.488930	0.472371	0.728436	0.296751	0.480508
2	DTC_optimised1	Imbalanced	0.600962	0.445633	0.773934	0.368873	0.511771
3	DTC_optimised2	Imbalanced	0.564583	0.483066	0.763507	0.364945	0.520653
4	decision_tree2_best	Imbalanced	0.626667	0.502674	0.788152	0.420768	0.557864
10	decision_tree2_best	Balanced	0.461538	0.545455	0.709953	0.297720	0.500000
11	decision_tree_entropy_bal	Balanced	0.458571	0.572193	0.706635	0.303534	0.509120
12	decision_tree3_bal	Balanced	0.474900	0.843137	0.710427	0.405280	0.607579
13	decision_tree4_bal	Balanced	0.564583	0.483066	0.763507	0.364945	0.520653
14	decision_tree2_best_bal	Balanced	0.451844	0.786096	0.689573	0.356587	0.573845

1. As compared to a 0.74 accuracy in balanced data in previous algorithm, here we were able to enhance the accuracy to about 0.763.
2. But again a model with good accuracy and a good recall score is not achieved in decision tree modelling.
3. Recall of 0.84 is achieved but with lower accuracy score (0.71).

RANDOM FOREST

	Model	Data Balanced/Imbalanced	Precision Score	Recall Score	Accuracy Score	Kappa Score	f1-score
5	RFC_model1	Imbalanced	0.606509	0.365419	0.768246	0.320141	0.456062
15	RFC_model2	Balanced	0.570397	0.563280	0.771090	0.411269	0.566816

We were keen on getting a better classification so we did try some ensemble method. Random Forest being our first choice for the same

1. We further were able to enhance the accuracy to 0.77.
2. But recall score deteriorated further.

BOOSTING

	Model	Data Balanced/Imbalanced	Precision Score	Recall Score	Accuracy Score	Kappa Score	f1-score
6	adaboost_model	Imbalanced	0.662651	0.490196	0.798104	0.436002	0.563525
7	Gboost_model	Imbalanced	0.684211	0.463458	0.800474	0.430260	0.552604
8	XGB_model	Imbalanced	0.610577	0.452763	0.777725	0.379458	0.519959
9	XGB_model_best	Imbalanced	0.625571	0.488414	0.786256	0.411298	0.548549
16	adaboost_model_bal	Balanced	0.499470	0.839572	0.733649	0.439433	0.626330
17	Gboost_model_bal	Balanced	0.502151	0.832442	0.736019	0.441026	0.626425
18	XGB_model_bal	Balanced	0.545736	0.627451	0.762085	0.418321	0.583748
19	XGB_model_bal_best	Balanced	0.521898	0.764706	0.751185	0.444972	0.620390

Why Recall score is important here?

As it determines how well our model classifies the customers who are more likely to leave.

1. Boosting techniques like ADABOOST, GBOOST and XGBOOST are used here for classification.
2. The best model are highlighted.

greatlearning

SOME MORE ALGORITHMS....

	Model	Data Balanced/Imbalanced	Precision Score	Recall Score	Accuracy Score	Kappa Score	f1-score
20	gnb_model	Imbalanced	0.515738	0.759358	0.746445	0.435518	0.614275
21	gnb_bal_model	Balanced	0.483940	0.805704	0.719905	0.408019	0.604682
22	sv_model	Imbalanced	0.665796	0.454545	0.794313	0.413782	0.540254
23	sv_bal_model	Balanced	0.498335	0.800357	0.732701	0.426182	0.614227
24	Lgbm_model	Imbalanced	0.635934	0.479501	0.788626	0.412438	0.546748
25	Lgbm_model_bal	Balanced	0.544863	0.746881	0.766825	0.465845	0.630075

1. Different models applied here are Naïve Bayes, Support vector machine and Light gradient Boosting Model.
2. For the balanced data a good accuracy is required along with the good recall score. This is achieved in the LGBM balanced model as highlighted.
3. Surprisingly this model gives us the kappa score which is best among all the models fitted earlier.

CONCLUSION

1. We started by cleaning the data and analysing it with visualization. Then, to be able to build a machine learning model, we transformed the categorical data into numeric variables (feature engineering). After transforming the data, we tried different machine learning algorithms using default parameters.
2. Finally, we concluded that the accuracy that we are getting in applying different algorithms is near about 0.77.
3. The best set of metrics achieved is in LGBM model with balanced data. Having good accuracy score, recall as well as kappa score.
4. This accuracy is subjected to the fact that there is no over-fitting or under-fitting.
5. So, this model can be deployed in the classification of customers. That is they will remain associated with the company or terminate the services.
6. Balancing of data was important here because then we were confident over the accuracy achieved.
7. Our aim was to build a model centered around the data provided to accurately classify the customers which are likely to terminate the services soon. This model can achieve the objective with 77% confidence.
8. The data provided is for certain period of time. If more observations are added with longer time period then surely there will be more information on churn rate.

FINAL RECOMMENDATIONS

On the basis of the EDA and various model built after data analysis following recommendations are given to the telecommunication company:

1. New customers are likely to leave company early so try to retain the new customers with lower charges and better services.
2. Introduction of packages for whole family is much needed so that the customer gets the value for money.
3. Company is not able to gain trust of new customers but once a customer is associated with the company for more time then for sure it remains with the company for a longer period.
4. Model which is built can predict the customers which are likely to churn out with an accuracy of 77%. So it can be used by the PR team to bring some perks and packages for them in order to retain them.
5. Considering the fact that customers using fibre optics are more likely to terminate the services. This is may be due to the fact that service is liked by people but it is not economical. So economical plans with better service is must to retain the customers.

Thank you...