2022

Case: Churn Prediction Modelling for a Telecom Company

REPORT FOR AN END-TO-END MACHINE LEARNING FLOW

IRISH MEHTA

Table of Contents

1.) Exploratory Data Analysis

1.1) Assumptions:

1) Scope of data: Before 31st January 2020
   Reason: Pre-pandemic data is considered here because a lot of factors have come into play post that period, mainly the idea of Remote work and popularity of streaming services. It would not make sense to consider data post the spread of the pandemic.

2) Definition of Churn: As per the documentation, churn is defined as all the people that did not use the services in this month (or last month if we are looking at 1st of the current month). Based on assumption 1, this would mean that all the customers that have not paid for the service after 31st December 2019 are churned customers.

3) Because the data spans over 6 years, it has to be assumed that the services and external factors have stayed constant in the past 6 years and there is no change in the services / pricing / availability / workforce

1.2) Initial constraints of data points:

i) Customer ID is unique and each row pertain to one specifc customer ID
ii) Tenure has to be > 0 based on the definition of churn
iii) Monthly charges cannot be greater than total charges
iv) A user cannot opt for any of the subscriptions if internet services is not taken
v) Multiple lines are not possible without enablement of Phone service
vi) Gender is Male/Female and does not include non-binary genders
vii) There are 3 types of contracts, it is possible that the user has churned out during their contract or they have churned out after their contract is over

1.3) Addition of Time Interval to the Data

- The baseline time interval added to all the customers is the time of their registration/sign-up. This information is derived from the given tenure of all customers and assumption 1 (stating the scope of the data). Based on this distribution, availability of parameters such as registration year and registration month is ensured, hence allowing the presence of time-oriented features for the model.

1.4) Missing Value Handling:

- Majority of the dataset is cleaned except the presence of 11 missing entries in TotalCharges column. One further inspection, it seems these 11 rows also do not follow one of the constraints (Tenure > 0 months) and that means this data cannot be trusted. These values need to be removed because they seem to be having incorrect data. 0.15% of the data has been removed.

1.5) Outlier Handling:

- Techniques such as Z-score method or InterQuartile range help in the case of Normal distributions, but most of the variables in the given data are not Normal. Hence a simple outlier removal on the basis of Percentiles (Winsorization method) is used.

1.6) Correlations:

1. Subscription based services (Internet services, Online Security, Online Backup, Device Protection, TechSupport, Streaming TV and Steaming Movies) are highly correlated to each other. Plausible reason: Generally, customers prefer taking multiple services from the same company for many reasons like convenience, compatibility, uniformity etc. Certain patterns are clearly visible, for example:

    i) The correlation of Internet Service with all the other subscription services. People who subscribe for internet would naturally want to utilize the most that they can get out of that service, so they could opt for Online Backup, Streaming TV/Movies, Online Security, etc.
    ii) People who opt for Online Backup would have a tendency to take Online Security/Device Protection for their backup to be safe.
    iii) Customers opting for multiple services would have frequent interactions with the Technical support team for small problems and hiccups, hence Technical support is highly correlated to the other services as well

2. Dependents is significantly correlated to Partner, which signifies that a lot of people that have partners as dependents, and could also signify that a lot of people having dependents also have partners.
3. Contract seems to be negatively correlated to Churn, which would imply that shorter contracts correspond to higher churn
4. Variables that are significantly correlated to churn are Online Security, Device Protection, Tech Support, Payment Method
5. Phone Service is highly correlated to Multiple Lines, implying that most of the customers that have a phone service subscription have it in the form of multiple phone lines and not a single one

1.7) Distributions:

1) Tenure: Bimodal distribution having two peaks of similar strength
2) Monthly Charges: Bimodal distribution again, with significantly different peak heights
3) Total Charges: Skewed distribution. Positive and Right skewed.
4) Gender: 50.4% M and 49.5% F. Almost similar counts.
5) Senior Citizen: 16% senior citizens, 84% non-senior citizens.
6) Dependents: 70% don’t have any dependents and 30% have dependents
7) Partners: 51% don’t have a partner and 49% have a partner
8) Contract: 55% of the customers have month-to-month contract, whereas the remaining have opted for yearly or two yearly subscriptions
9) Phone Service: 90% of the people have a phone service. Remaining 10% do not have an active phone service
10) Multiple phone lines: 59% of phone service subscribers have a multi-line service. Remaining have opted for a single line service
11) Internet Service: 20% of the people have not opted for internet service. 45% have opted for Fibre optic service and 35% have opted for DSL connection
12) Streaming TV & Streaming Movies have a 60% subscribe rate of internet enabled customers.
13) Online backup, Device Protection and Online Security have a 68-70% subscribe rate from internet enabled customers.
14) Tech support also has a 70% subscribe rate from internet enabled customers.
15) 60% customers have opted for paperless billing.
16) 35% of the customers use Electronic check for payment and 75% of the people either use Mailed Check, bank check or credit card

17) Registration Year: Customers are either extremely old or extremely new. This also follows a sort of two peak distribution

1.8) Hypothesis Testing

1.8.1) Hypothesis 1: Relation of Tenure & Churn

Null Hypothesis: *There is no relation between the tenure of the user and their churn*
Alternate Hypothesis: *Tenure of a customer is related to his/her churn*

- Variable to be tested for the hypothesis: Tenure
- Distribution of variable to be tested for the hypothesis: Bimodal

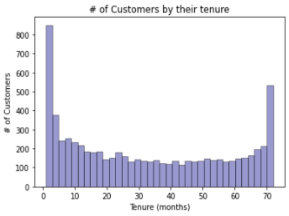


Figure 1 Distribution of Tenure

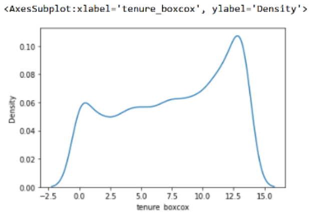

```
<AxesSubplot:xlabel='tenure_boxcox', ylabel='Density'>
```

Figure 2 Distribution of boxcox transformation of Tenure

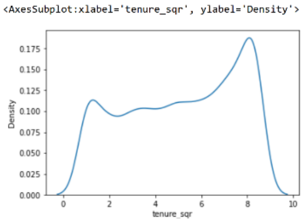

```
<AxesSubplot:xlabel='tenure_sqr', ylabel='Density'>
```

Figure 31 Distribution of square root transformation of Tenure

Assumptions and Feasibility

1. Considering an error threshold (Alpha) of 0.05 After grouping the customers into two groups, one that churned out and one that did not Following are the results of multiple statistical tests to check how significantly different is the mean of tenure of both the groups
2. T test has been used to determine the validity of the hypothesis, but because the input variable is not normally distributed, there is a high chance that this data does not represent the true relation between the means of the two groups.
3. Other statistical tests have not been used because comparison is being done only between 2 groups
4. Transformations such as box-cox, logarithmic transform, square root transform to reduce positive skew were tried but it did not give a normal distribution in any case hence the original bimodal distribution is considered

Statistical Evidence

- **Statistical Test**: T-test (one way):
- **Test statistic**: T-statistic. Helps to explain the difference in means of both groups. Value is 31
- P-value: 9.4 X 10^{-207}
- **Result**: In the T- test, we see that the P value is way below **0.001** and the statistic measure in each test is extremely large implying that the means of both the group are significantly different. Based on this statistical evidence, we can safely reject the null hypothesis and favor the alternate hypothesis, stating that **The tenure of customers significantly affects the churn of the customers**

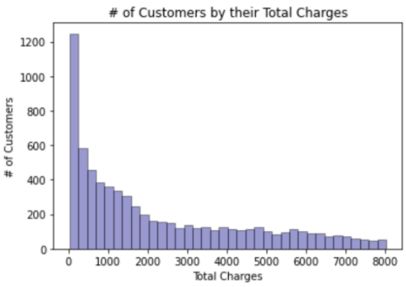
1.8.2) Hypothesis 2: Relation of Total Charges & Churn

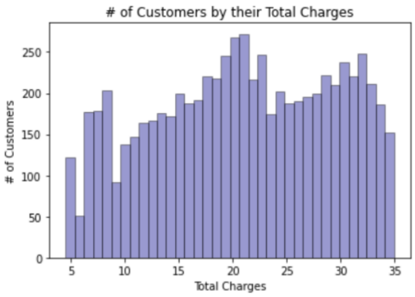
Null Hypothesis: There is no relation between the total charges given by a user and his/her churn
Alternate Hypothesis: Total charges given by a user is related to his/her churn

- Variable to be tested for the hypothesis: TotalCharges
- Distribution of variable to be tested for the hypothesis: Right skewed (Positive) distribution
- Transformation function applied= Box cox transform

Assumptions and Feasibility

1. Considering an error threshold (Alpha) of 0.05 After grouping the customers into two groups, one that churned out and one that did not Following are the results of multiple statistical tests to check how significantly different is the mean of tenure of both the groups
2. T test has been used to determine the validity of the hypothesis, but because the input variable is not normally distributed, there is a high chance that this data does not represent the true relation between the means of the two groups.
3. Other statistical tests have not been used because comparison is being done only between 2 groups
4. The transformation function being used is box cox transformation because other transformations (square root, log, cube root) were giving distributions with a higher negative skew. Square root transform gives a distribution closer to Uniform distribution rather than Normal but it was the best result of all others





Statistical Evidence

- **Statistical Test**: T-test (one way):
- **Test statistic**: T-statistic. Helps to explain the difference in means of both groups. Value is 20
- P-value: 1.5 X 10^{-81}
- Result: In the T- test, we see that the P value is way below **0.001** and the statistic measure in each test is extremely large implying that the means of both the group are significantly different. Based on this statistical evidence, we can safely reject the null hypothesis and favor the alternate hypothesis, stating that **The total charges given by customers significantly affects the churn of the customers**

2.) Feature Engineering:

2.1) Raw Features

No.	Feature Name	Type	Description
1	Customer ID	Object	Identity for each customer
2	Gender	Object	Flag for gender of a customer
3	Senior Citizen	Int64	Flag for whether a person is a senior citizen (1) or not (0)
4	Partner	Object	Whether the customer has a partner or not
5	Dependents	Object	Whether the customer has dependents or not
6	Tenure	Int64	Number of months the customer has been with the company
7	Phone Service	Object	Whether the customer has a phone service or not
8	Multiple Lines	Object	Whether the customer has multiple phone lines
9	Internet Service	Object	Internet service provider of the customer
10	Online Security	Object	Flag for subscription of online security
11	Online Backup	Object	Flag for subscription of Online Backup
12	Device Protection	Object	Flag for subscription of Device Protection
13	Tech Support	Object	Flag for subscription of tech support
14	Streaming TV	Object	Flag for subscription of Streaming TV
15	Streaming Movies	Object	Flag for subscription of Streaming Movies
16	Contract	Object	Contract term of the customer
17	Paperless Billing	Object	Flag for whether the person has paperless billing
18	Payment Method	Object	Payment method that the customer has opted for
19	Monthly Charges	Float64	The amount charged to the customer monthly
20	Total Charges	Float64	Total amount charged to the customer
21	Churn	Object	Whether the customer churned or not

2.2) Derived Features

No.	Feature Name	Type	Description	Constraints
1	Registration Time	Datetime	The date of when a customer registered	-
2	Registration Year	Object	The year of when a customer registered	4 digits/characters
3	Registration Month	Int64	The month of when a customer registered	1 ≤ x ≤ 12
4	Total Payment Cycles	Int64	Number of payment cycles (count) that have passed since their registration based on their current contract. Eg. If a customer registered 30 months back and is currently on a Monthly subscription, the total payment cycles would be 30	For a given customer, total payment cycles cannot be > the tenure of that customer
5	Security_subscription	Int64	A flag signifying that the customer has subscribed to any one of the security service (Device protection or Online Security)	x ∈ [0,1]
6	Streaming Subscription	Int64	A flag signifying whether the customer has subscribed to any one of the streaming services (Streaming Movies or Streaming TV)	x ∈ [0,1]
7	Known_network	Int64	A flag signifying whether a customer has any Dependents/Partners	x ∈ [0,1]
8	Payment type	Object	A binary flag giving information about the type of payment method used by the customer. It takes two values, Automatic and Manual	x ∈ [automatic,manual]
9	Internet with Tech support	Int64	A flag signifying whether a customer has taken internet services with tech support or not	x ∈ [0,1]
10	Total internet based subscriptions	Int64	A sum signifying the total subscriptions taken by the customer	0 ≤ x ≤ 7
11	Payment Per month	Float64	Total charges given by the customer per month (lifetime). Given by TotalCharges/tenure	x>0
12	Service change impact	Float64	Percentage difference between the total charges the customer would have given (tenure * monthly charge) and the charges that a customer has given (total charges)	No constraints. However, values <-1 and > 1 should be extremely improbable

2.3) Explanation of Derived features:

The original dataset consists of multiple fragmented features that give information about the current state of the user and how engaged he is in the system.

However, the features lacked the following overall aspect for each customer which are tried to address by the introduction of derived parameters:

1) Quality:
    a. Registration Year
    b. Registration Month
    c. Known Network
    d. Total internet based subscriptions
2) Behavioural aspects (like payment habits, type of subscriptions
    a. Total Payment Cycles
    b. Security Subscription
    c. Streaming Subscription
    d. Payment Type
    e. Internet with Tech support
3) Impact aspects (like impact of cost change, service change, etc on any customer)
    a. Service change impact

3.) Predictive Modelling Pipeline:

3.1)Sampling Distribution for splitting Training set and Testing Set:

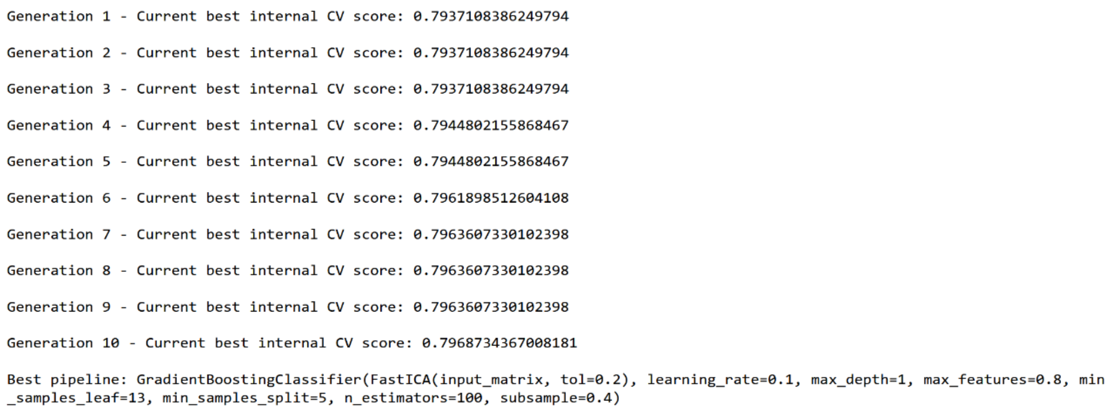
- **Method**: Stratified Sampling
- **Reason**: Stratified sampling is best when there is an inherent imbalance in the target variable. Here, the split for churn is 70% towards non churning customers and 30% towards churning customers.
- ***Note**: In real life scenarios, such a split in the dependent variable is considered to be more than good. An imbalance in the real world use case would have an imbalance of 95% and 5%.*

3.2) Best x Features:

- Method: Feature selection using Random Forest Classifier
- Reason: Feature selection using a Random Forest Classifier is extremely useful in generalizing the overall data and providing a highly accurate output. There are other selection methods like Principal Component Analysis but they are not as interpretable as a Tree based classifier is.
- Result: The top 6 features are *MonthlyCharges, payment_per_month, tenure,* *service_change_impact, total_payment_cycles, Contract_Month-to-month*

3.3) AutoML for Model Selection:

- Method: TPOT Classifier library for selection of the most optimum model
- Time consumed: 40mins
- Results:

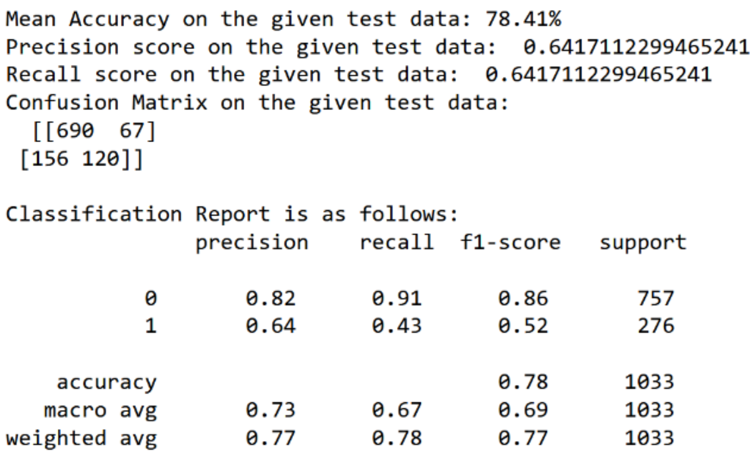


```
Generation 1 - Current best internal CV score: 0.7937108386249794

Generation 2 - Current best internal CV score: 0.7937108386249794

Generation 3 - Current best internal CV score: 0.7937108386249794

Generation 4 - Current best internal CV score: 0.7944802155868467

Generation 5 - Current best internal CV score: 0.7944802155868467

Generation 6 - Current best internal CV score: 0.7961898512604108

Generation 7 - Current best internal CV score: 0.7963607330102398

Generation 8 - Current best internal CV score: 0.7963607330102398

Generation 9 - Current best internal CV score: 0.7963607330102398

Generation 10 - Current best internal CV score: 0.7968734367008181

Best pipeline: GradientBoostingClassifier(FastICA(input_matrix, tol=0.2), learning_rate=0.1, max_depth=1, max_features=0.8, min
_samples_leaf=13, min_samples_split=5, n_estimators=100, subsample=0.4)
```

- Best Pipeline: Gradient Boosting Classifier
- **However, XGBoost classifier will be the selected pipeline because of it’s benefits in speed, regularization and flexibility.**

3.4) Model Training

- Input Data Points: 6 most important features, namely *MonthlyCharges*, *payment_per_month*, *tenure*, *service_change_impact*, *total_payment_cycles*, *Contract_Month-to-month*
- Sampling Method: Stratified Sampling
- Algorithm: XG Boost Classification
- Classification Profile:

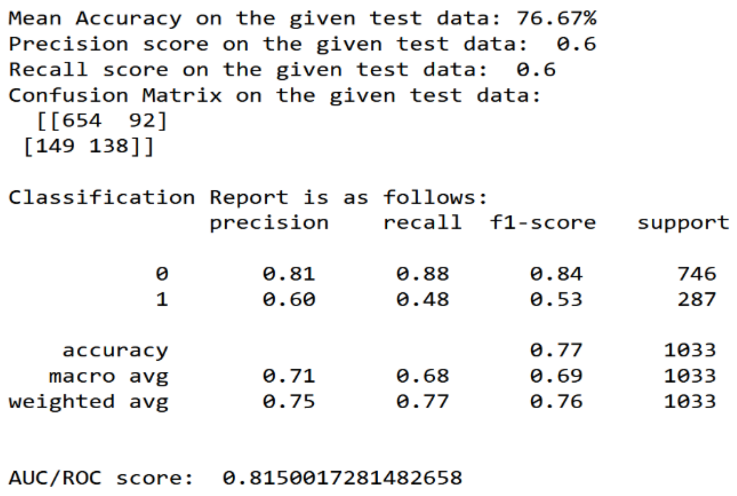


```
Mean Accuracy on the given test data: 78.41%
Precision score on the given test data:  0.6417112299465241
Recall score on the given test data:  0.6417112299465241
Confusion Matrix on the given test data:
  [[690  67]
 [156 120]]

Classification Report is as follows:
              precision    recall  f1-score   support

           0       0.82      0.91      0.86       757
           1       0.64      0.43      0.52       276

    accuracy                           0.78      1033
   macro avg       0.73      0.67      0.69      1033
weighted avg       0.77      0.78      0.77      1033
```

- Cross validation Method: Stratifiedkf
- Average Cross validation Score = 0.7924

3.5) Hyperparameter Tuning

- Method: GridSearchCV
- Updated classification report after hyperparameter tuning:



```
Mean Accuracy on the given test data: 76.67%
Precision score on the given test data:  0.6
Recall score on the given test data:  0.6
Confusion Matrix on the given test data:
  [[654  92]
 [149 138]]

Classification Report is as follows:
               precision    recall  f1-score   support

           0       0.81      0.88      0.84       746
           1       0.60      0.48      0.53       287

    accuracy                           0.77      1033
   macro avg       0.71      0.68      0.69      1033
weighted avg       0.75      0.77      0.76      1033


AUC/ROC score:  0.8150017281482658
```

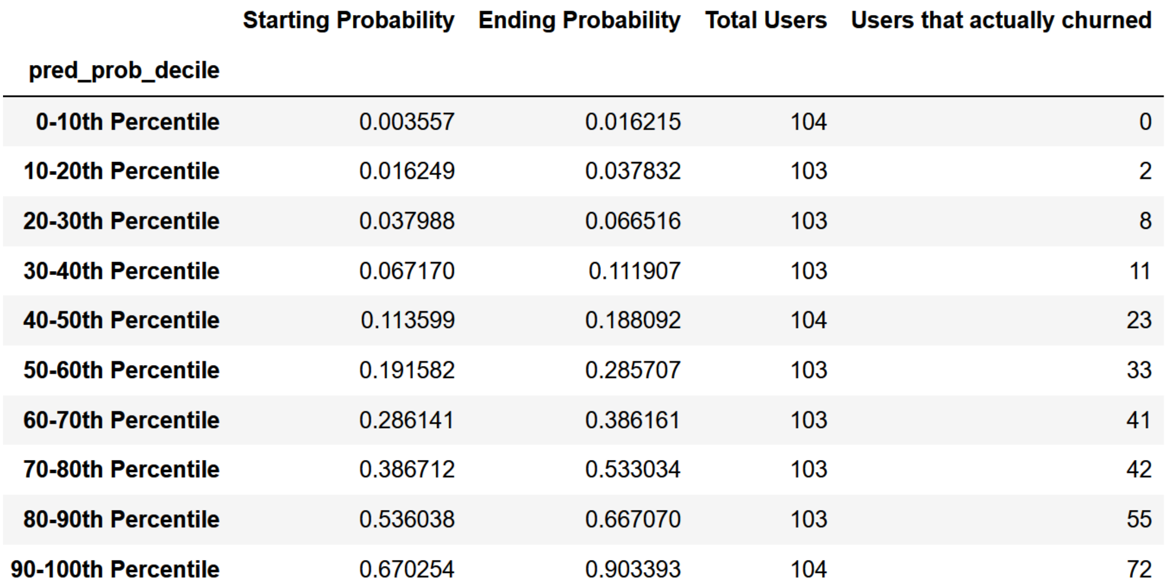
- Average Cross validation score= 0.7855

3.6) Interpretability and Explanation of Model output:

- We will be analyzing the result of the model, validate that result and try to understand the reason for the results.
- *Note: Precision helps in validating the correctness of the positive predictions, i.e it helps to understand how many of our predictions of churning customers actually turned out to be true. Recall on the other hand measures the ability of the model to figure out the true positives, i.e it measures how good our model is in detecting Churn.*
- **For our business use case, we would like to increase the recall and try to capture maximum of those customers that will churn. This is because we don't want to miss out on customers that have the potential to churn.**

Prediction Summary

- Total samples on the Test data: 1033



	Starting Probability	Ending Probability	Total Users	Users that actually churned
pred_prob_decile				
0-10th Percentile	0.003557	0.016215	104	0
10-20th Percentile	0.016249	0.037832	103	2
20-30th Percentile	0.037988	0.066516	103	8
30-40th Percentile	0.067170	0.111907	103	11
40-50th Percentile	0.113599	0.188092	104	23
50-60th Percentile	0.191582	0.285707	103	33
60-70th Percentile	0.286141	0.386161	103	41
70-80th Percentile	0.386712	0.533034	103	42
80-90th Percentile	0.536038	0.667070	103	55
90-100th Percentile	0.670254	0.903393	104	72

***Here probabilities are the predicted probabilities given by the model. This tables essentially shows the distribution of predicted probabilities and allows us to figure out the ideal threshold for our use case

Test Metric:

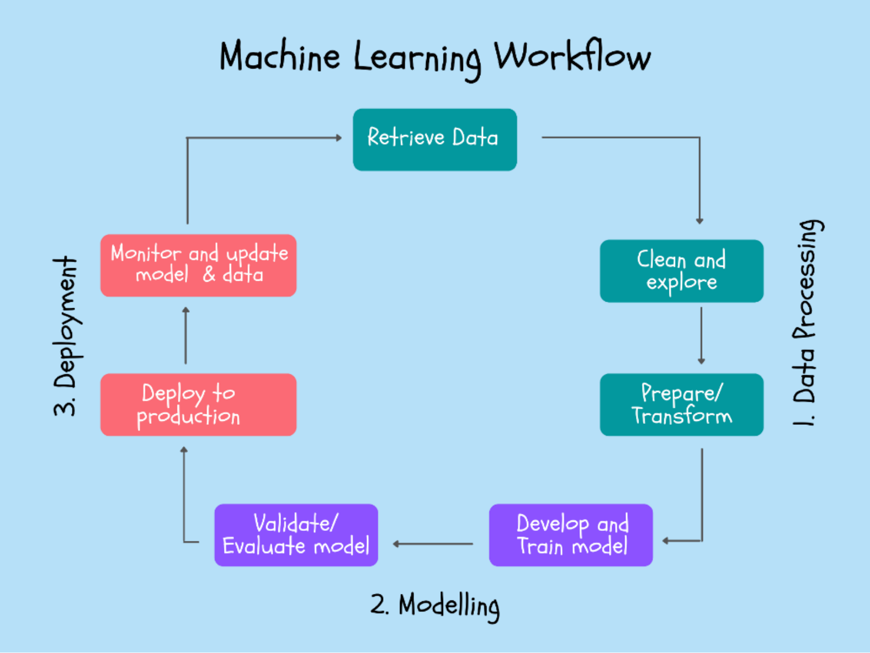
Since the selected test metric is Recall, following is the change in recall as the probability threshold changes.

Probability Threshold	Recall value
0.6	64%
0.7	76%
0.8	86%

- **We can go ahead and deploy the model at a threshold of 0.8, below which a customer not churn and above which the customer will churn.**

4.) Model Deployment Pipeline:

A machine learning pipeline is a way to control and automate the workflow it takes to produce a machine learning model. Machine learning pipelines consist of multiple sequential steps that do everything from data extraction and preprocessing to model training and deployment.



In the funnel till now, Data Processing and Modelling have been completed. The final phase of an ML pipeline is the deployment of the trained model. Deployment is done in two steps

1) Deploying a Machine Learning Model
2) Monitor, Log, Update the Model and its performance

4.1) Deploying the Machine Learning Model:

- Deployment of an ML-model simply means the integration of the model into an existing production environment which can take in an input and return an output that can be used in making practical business decisions.
- In the current use-case, the predictive model regarding churn of customers of a Telecom company has to be deployed in production such that the output of this ML model can be utilized.

Before discussing about the process, please go through the following assumptions:

1) Churn prediction happens for all the customers at a specific given time in their journey with the company.
   Here, multiple approaches are possible, for example, churn prediction can be done for all the customers (new and old registrations) at the start of every month and it can be predicted whether they will churn this month or not. Here however, the registration date of all customers is different so the prediction cycles for all of them would also be different. Another approach is predicting all the customers that will churn out during this quarter or not.

   We will be going with the assumptions that churn will be calculated monthly for all the customers based on when their payment cycle is.

2) To make the deployment more efficient and operationally feasible, following are the segments we can focus on:
    a. Modularity:
        i. The entire pipeline, starting from EDA and ending at Model training, should be fragmented into code pieces and each step should be well documented. This would ensure that the model can be updated as and when required
    b. Reproducibility:
        i. The pipeline should be developed such that when someone tests the processes later on, the outputs should be the exact same
    c. Scalability:
        i. Based on the use case and business understanding, sufficient processing, load handling and latency should be accounted for.
        ii. This can be done by creating multiple processes while deployment as well as ensuring that the latency is reduced by optimizing the flow of the development pipeline

4.1.1) Machine Learning Deployment Architecture

There are 4 possible architectures:

- **Train by batch, predict on the fly and interact with web interface**: training and persisting are done offline while prediction is done in real-time.
- **Train by batch, predict by batch and interact with shared database**: training and persisting are done offline while predictions are done in a distributed queue which is almost similar to a real-time prediction
- **Train, predict by streaming**: both training and predicting are done on different but connected streams.
- **Train by batch, predict on mobile(or other clients)**: similar to type 1 but the prediction is made on customer gadget.

For churn prediction of telecom customers, we will train data by batch and predict it on the fly via any web interface/API. Following is the flow of the architecture:

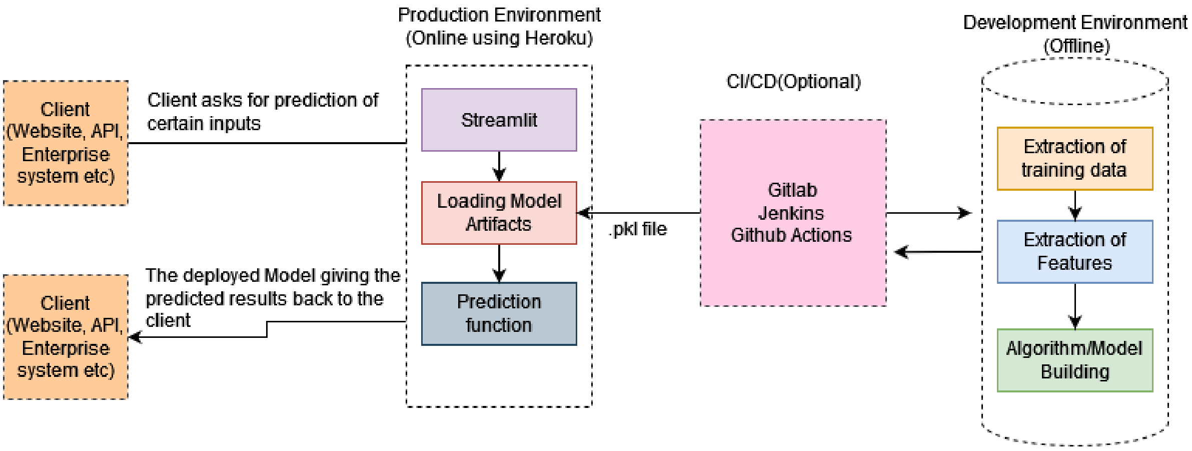


Figure The selected Architecture for this Modelling

4.2) Post Deployment Processes:

4.2.1) Monitoring

- Data Quality and Integrity: There are a lot of processes when we transform data. If the quality is compromised at any stage, the model will break. Similarly, a mismatch in the data integrity can also break the pipeline. For example, a case when there arises a problem in the database system of our telecom company and a lot of parameters are showing missing values. In this case the model deployment cycle will break.
- Model Validation: It needs to be monitored whether the deployed model is actually solving the problem it was intended to. Additionally, a concept of data drift can also occur here. For example, The customers of our telecom company can change over decades, so there has to be an assessment of how the population is behaving year after year. Techniques such as measurement of PSI can be used here.
- Load balancing and Costs Measurement: Based on the available resources and budget, input load, memory utilization, latency, costs, etc need to be monitored at all stages

4.2.2) Logging

- Root Cause Analysis: In cases where the monitoring pipeline fails, the entire flow uptil the point of failure needs to be logged so as to understand the reason and cause of failure. This would potentially help in troubleshooting the problem and get the system up and running quicker

4.2.3) Iteration/Updation

- Improvement in prediction accuracy: Based on model prediction logs, we can iterate over the predicted data and retrain the model on newer data. This iterative procedure ensures generalization of the data and also allows the model to learn better. We can lay out a plan to update this churn prediction model every 6 months and ensure that the model performance is up to the mark

5.) Demonstration of Deployed Model

- Web App Framework used: Streamlit
- Production Environment created through Heroku
- Link of the model web page: https://poc-telco-churn.herokuapp.com/

6.) Experimentation flow after Roll-out

Post the model rollout, there will be a monthly prediction of the probability of churn of a customer. Based on this intel, we can run the following experiments:

1. We create a target group of all the people that have a high likelihood of churn and try to give them cheaper deals on longer contracts. Control group would be to gauge the baseline
2. We create a target group of customers that have a high likelihood of churn and try to incentivise them by giving free subscriptions of online backup/security/tech support/streaming tv or movies etc. The control group won’t be given any such incentive.
3. We create a target group having a set of users with low likelihood of churn (or prediction class=0) and try to cross-sell them. This means that we tell them to take phone lines alongwith internet line or vice versa, take streaming services with security services etc. Cross selling would be really great to increase the engagement of the user. The control group can be used to set a baseline
4. We create a target group having a set of users with low likelihood of churn (or prediction class=0) and try to upsell them. If they have a lower tier of Internet service, we can incentivise to tell them to buy a higher one at some discount. This would help the company generate more revenue