Solution Architecture: -Low Level Design

Telecom Customer Churn

| **Written By** | Devanshu Shrivastava |
| --- | --- |
| **Document Version** | 1.0 |
| **Last Revised Date** | 02-Jan-2022 |

**Document Control**

### Change Record:

| **Version** | **Date** | **Author** | **Comments** |
| --- | --- | --- | --- |
| 1.0 | 02-Jan-2022 | Devanshu Shrivastava | Draft Version of Document. |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

### Reviews:

| **Version** | **Date** | **Reviewer** | **Comments** |
| --- | --- | --- | --- |
|  |  |  |  |

Contents

1. [**Introduction**](#_gjdgxs) 1
   1. [**What is Low-Level design document?**](#_30j0zll) 1
   2. [**Scope**](#_1fob9te) 1
2. [**Architecture**](#_3znysh7) 2
3. [**Architecture Description**](#_2et92p0) 3
   1. **Acquire** [**Data**](#_tyjcwt) 3
   2. [**Data Transformation/Cleanup**](#_3dy6vkm) 3
   3. **Feature Engineering and Selection**3
   4. **Feature Engineering and Selection**3
   5. [**Model Building and Evaluation**](#_17dp8vu) 4
   6. **Model** [**Deployment**](#_26in1rg) 4

# Introduction

## What is Low-Level design document?

The goal of LLD or a low-level design document (LLDD) is to give the internal logical design of the actual implementation of Telecom Customer Churn Use case. LLD describes the class diagrams with methods and relations between classes and program specs. It describes the modules so that the programmer can directly code the program from the document.

## Scope

Low-level design (LLD) is a component-level design process that follows a step-by-step [refinement](https://en.wikipedia.org/wiki/Refinement_(computing)) process. This process can be used for designing data structures, required software architecture, source code and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work

# Architecture



# Architecture Description

## Acquire Data

Provided dataset contains 7043 records for sample population data of telco customer acquired from the service desk as part of input to the Use Case that needs to be designed. Dataset contains 21 attributes out of which last attribute “Churn” is the target attribute. Each attribute is described as below.

Each row represents a customer, each column contains customer’s attributes described on the column Metadata.

The data set includes information about:

* Customers who left within the last month – the column is called Churn
* Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
* Customer account information – how long they’ve been a customer, contract, payment method, paperless billing, monthly charges, and total charges
* Demographic info about customers – gender, age range, and if they have partners and dependents

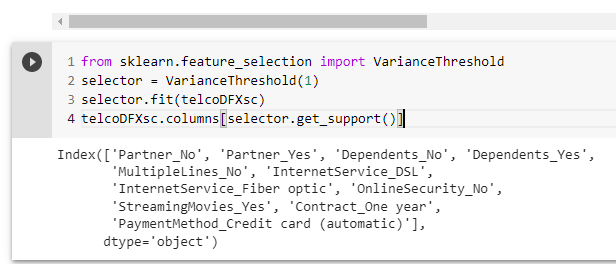
## Data Transformation/Cleanup

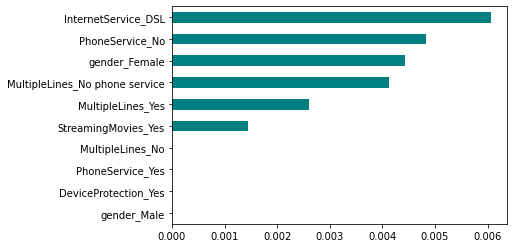
In the Transformation Process, we will convert data type of few fields like Total Charges to float as it has been derived to be object by panda’s interpreter. Also, Total Charges column has few records with blank space we shall replace the same with NaN and drop these records further from our dataset before applying in model. This decision has been taken after experimenting with imputation of this records by recalculating values using formulae tenure \* Monthly Charges which does not improve the outcome of Model. When these records are dropped Model provides better results.

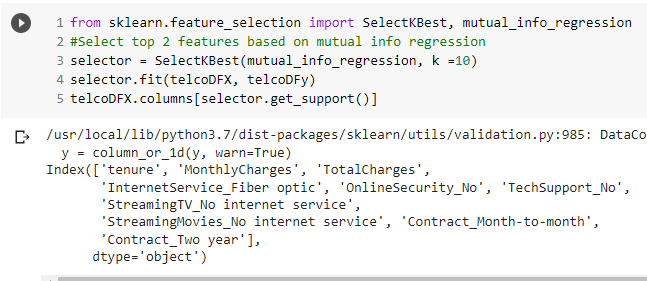
After removing the customer ID column as it provides unique identity to each record and thus denies variation in machine learning. It is observed that we have couple of duplicate records dropping this record was tested and it was observed it had no significant impact on model outcome thus data was kept as is and duplicates were not dropped.

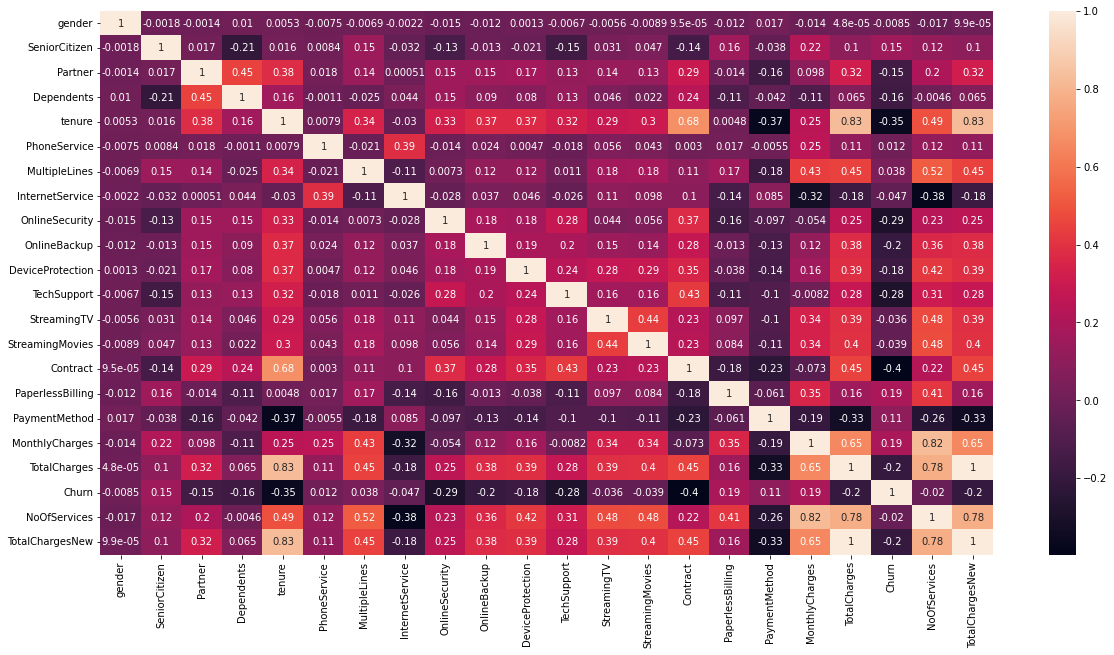
## Feature Engineering and Selection

This step includes applying methods to identify best n features which can be further used to improve model performance and weed out features which do not have significant impact or negatively impact the model performance. We have used methods like Variance Threshold, select K Best features, Information Gain and Correlation Heatmap to identify features with high variance or features which are highly correlated in which case only 1 of the correlated features will be used in the final model.



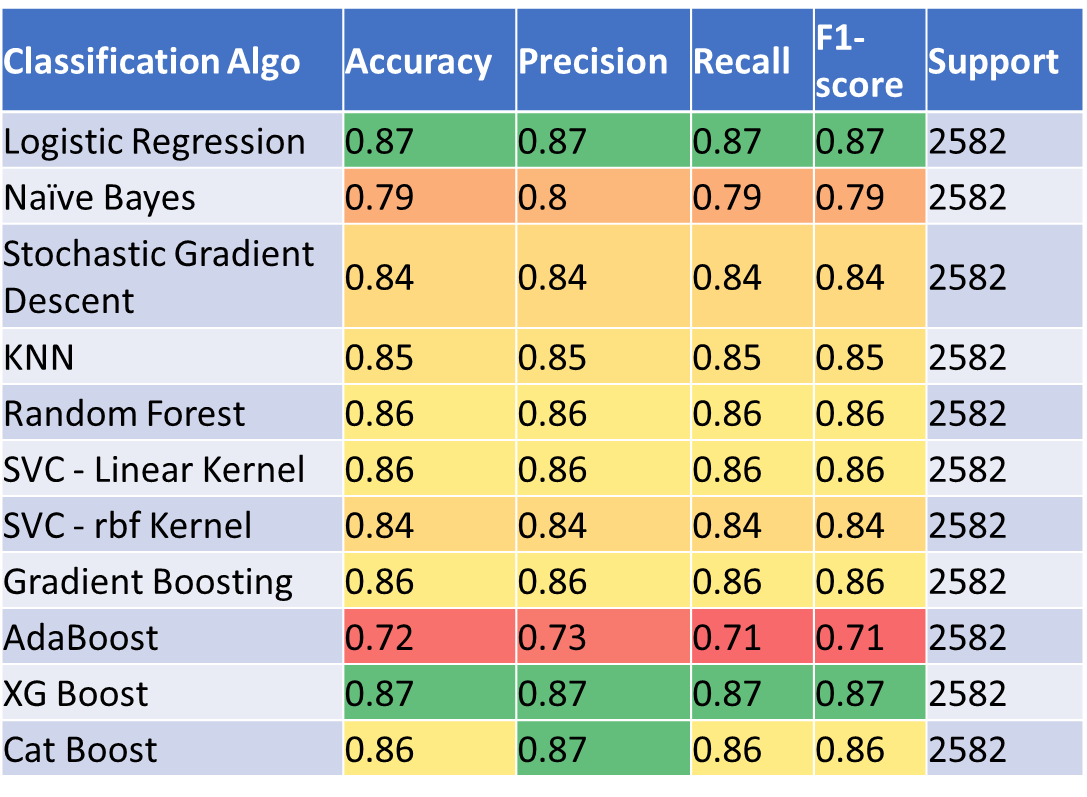




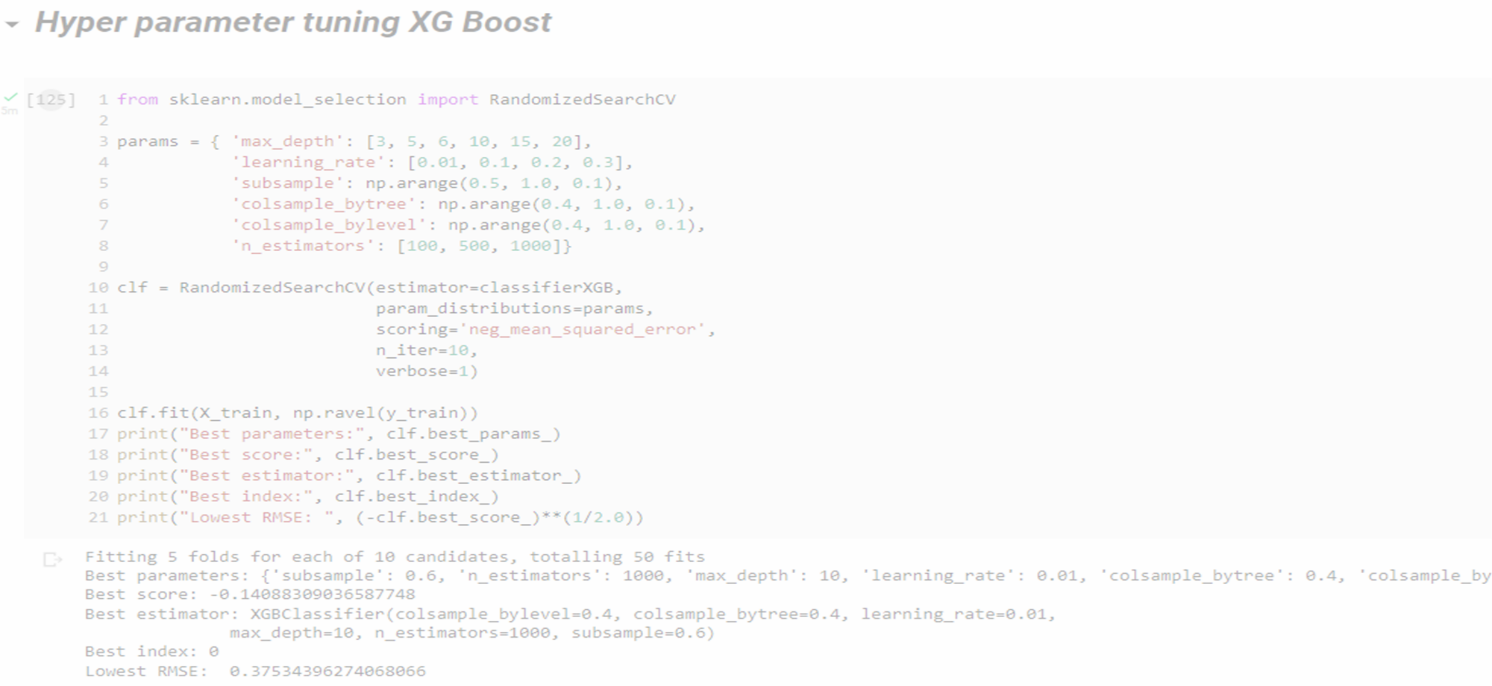


## Model Building and Evaluation

We Build and Evaluated 11 Classification Models with mentioned feature selections, data preprocessing, Standardization and Normalization mentioned in slides above. Best performing model is XG boost with accuracy of around 87.45 % while Linear Regression performed accuracy of 86.9 %.



Applied RandomSearchCV to best performing model XG Boost after trying multiple iterations and hyper parameter combinations it was observed that the default values used initially provided the best results. Simlary RandomSearchCV was applied to 2nd best performing model Logistic Regression after trying multiple iterations and hyper parameter combinations it was observed that the default values used initially provided the best results.



## Deployment

We will be deploying the model to customer premise/Cloud based on where the customer requirement. Real time test data will be received from customer service desk via a Rest API which will be preprocessed, scaled/Normalized, and then passed to the model for prediction. Resultant outcome will be displayed on a simple dashboard or intimated to agents via push notification with general recommendations so action can be taken proactively so as to stop or reduce on Churn rate.

This is a workflow diagram for the Telco Customer Churn.

