**Freelancing Customer Segmentation**

# Technical Design Document

**Version 1.0**

RESTRICTED DISTRIBUTION

The information is standard Company Confidential, but due to its sensitivity, it has restricted distribution and viewing within iNeuron.

Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| **Date Issued** | **Version** | **Description** | **Author** |
| 3/11/2020 | 1.0 | Initial Draft |  |

Contributors

The content of this document has been authored with the combined input of the following group of key individuals.

|  |  |
| --- | --- |
| **Name** | **Section Worked Upon** |
|  | Initial Draft |

Document Classification

|  |  |
| --- | --- |
| **Classification** | Company Confidential |
| **Definition** | Information is Group confidential and needs to be protected |
| **Context** | Where the loss of information confidentiality would result in significant harm to the interests of the organisation, financial loss, embarrassment or loss of information |

Contents

[1 Technical Design Document 2](#_gjdgxs)

[1. Introduction](#_30j0zll) 4

[High level objectives 5](#_1fob9te)

[2 Workflow Overall 6](#_3znysh7)

[Application Flow 6](#_2et92p0)

[Exception Scenarios Overall](#_tyjcwt) 7

# Introduction

The goal here is to build an end to end automated Machine Learning solution where the user will only give the data and select the type of problem, and the result will be the best performing hyper tuned Machine Learning model. The user will also get privileges to choose the deployment options.

This project shall be delivered in two phases:

Phase 1: All the functionalities with PyPi packages.

Phase2: Integration of UI to all the functionalities.

The technical design document gives a design blueprint of the freelancing customer segmentation project. This document communicates the technical details of the solution proposed.

In addition, this document also captures the different workflows involved to build the solution, exceptions in the workflows and any assumptions that have been considered.

Once agreed as the basis for the building of the project, the flowchart and assumptions will be used as a platform from which the solution will be designed.

Changes to this business process may constitute a request for change and will be subject to the agreed agility program change procedures.

**Note: All the code will be written in python version 3.7 and above.**

As a part of customer segmentation, we need to define and create the customer segmentation based on main features which customer and recommendation teams has identified as a part of freelancer websites for example- Based on budget (High, Medium, Low), Location(Based on geographical location), Employees feedback, Employer feedback, Average time to finalize job etc.

## High level objectives

The high-level objectives are:

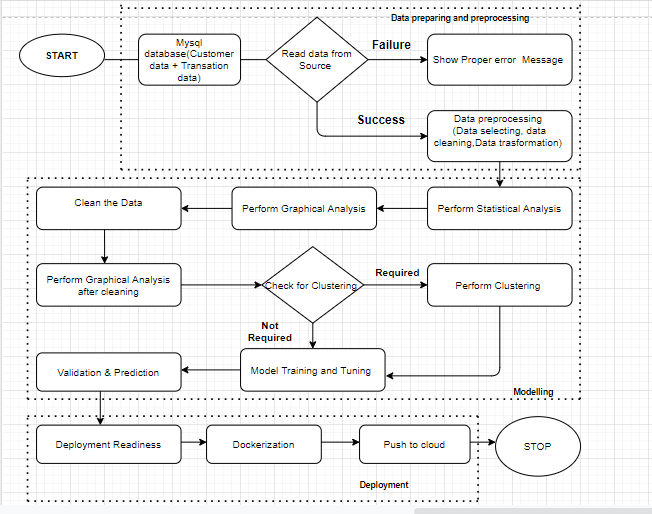
1. Enable reading/loading of data from the various sources and convert them into pandas data-frame(details mentioned in the Data Ingestion Section).
2. Enable reading various file formats and convert them into pandas data frame (details mentioned in the Data Ingestion Section).
3. Give user the option to specify feature and target columns.
4. Give user the option to select the problem type, viz. Regression, Classification (include anomaly detection), Clustering or Time Series.
5. Perform statistical analytics of the data and prepare a table for the analysis and show it on screen.
6. Perform graphical analysis for the data and Showcase the results (graphs) on the screen.
7. Perform data cleaning operation with all the steps required and showcase a report on screen.
8. After data cleaning showcase the graphical analysis once again for comparison.
9. Check whether clustering is required or not.
10. Choose the appropriate ML model for training.
11. Perform model Tuning.
12. Create a list of top 3 models and show multiple metrics for them.
13. Give option for prediction.
14. Give options for docker container creation.
15. Give option for automatic cloud deployment.

**Phase 1:** Create Pypi packages

**Phase 2:** Create UI

# Workflow Overall

# Application Flow



**Data preparing and Pre-processing** : Data pre-processing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.

**Modelling:** A machine learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.

* K-Means Clustering
* Density-Based Spatial Clustering of Applications with Noise
* Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM)
* Agglomerative Hierarchical Clustering
* OPTICS Clustering
* Affinity Propagation
* Birch Clustering

**Deployment**: Deployment is the method by which you integrate a machine learning model into an existing production environment to make practical business decisions based on data.

Some deployment platform Google Cloud Platform (GCP) is one of the primary options for cloud-based deployment of ML models, along with others such as AWS, Microsoft Azure,etc

- Training a machine learning model on a local system.

- Wrapping the inference logic into a flask application.

- Using docker to containerize the flask application.

- Hosting the docker container on an AWS ec2 instance and consuming the web-service.

## Exception Scenarios Overall

|  |  |  |
| --- | --- | --- |
| **Step** | **Exception** | **Mitigation** |
| User gives Wrong Data Source | Give proper error message | Ask the user to re-enter the details |
| User gives corrupted data | Give proper error message |  |
| If the cluster contains only one class | No error message required | Handle this exception internally. User doesn’t know. |
| Deployment credentials are wrong | Give proper error message | Ask for the details to be entered again |

# 