**High Level Design Document**

**Loan Acceptance**

****

**Project Guide:**

**Created By:**

## **Introduction**

Loans are the core business of banks. The main profit comes directly from the loan’s interest. The loan companies grant a loan after an intensive process of verification and validation. However, they still don’t have assurance if the applicant is able to repay the loan with no difficulties.

Online leading has gained popularity because of its higher efficiency in offering credit to consumers and small businesses. It is one of the industries that has been disrupted by technology with the electronic lending platform.

Loan application decision is made automatically with electronic data driven algorithms. Online lenders have the flexibility to offer small loans with short term maturities. Borrowers that are excluded from traditional banking systems can hence have the chance to access credit.

Online lending has evolved with platforms connecting lenders and borrowers to more diversified business models such as direct lending, financial institutional partnerships.

## **Problem Statement**

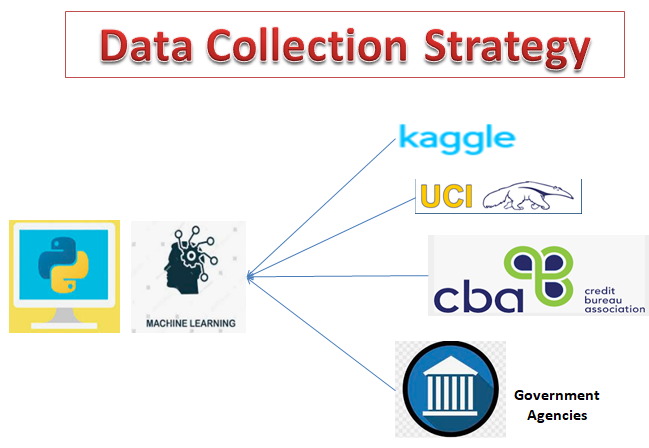
Build an automated system for loan application processing (decision to approve or reject the loan) based on the various parameters as decided by the machine learning algorithm.

The purpose of the document is to explain the High architecture that would be used for developing the automated loan acceptance system. The architecture diagram will provide an overview of an entire system, identifying the main components that would be developed for the product and their interfaces.

## **Objective**

The objective of this document is to present the brief overview of the technical architecture of the automated loan acceptance. It describes the complete journey -> Data Collection to the Best Model Selection to the Deployment of the ML model for end user’s usage.

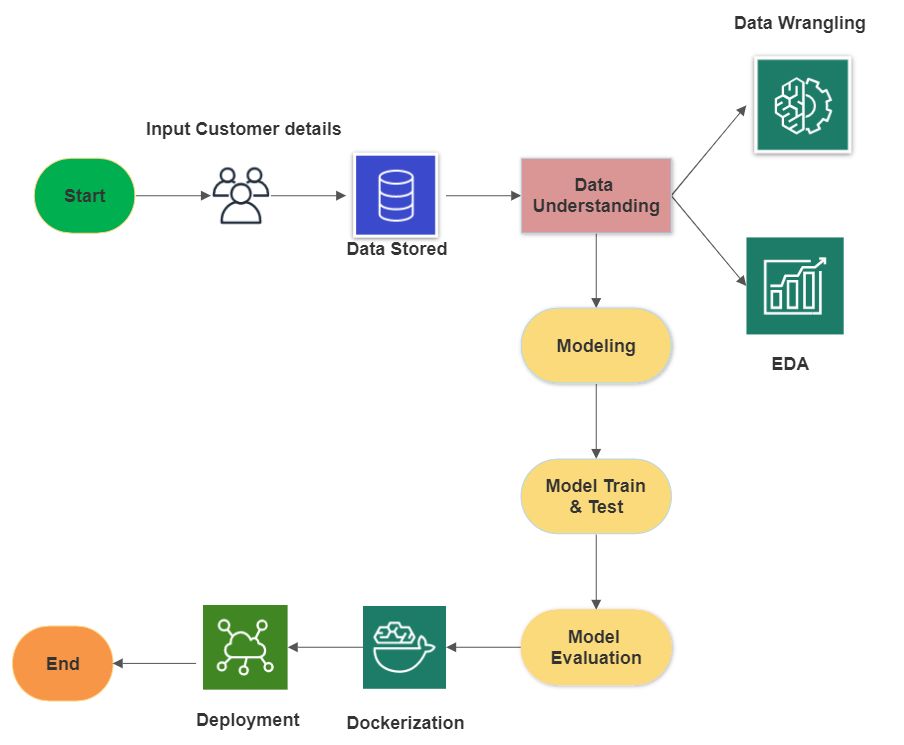
## **Data Collection Strategy**



For good performance of the model we need a reasonable quantity of data with sound quality. The data collected for training should be able to capture the variances of the population and should represent the population best way possible. The data used for the prediction should come from the same population as the one used for training. Some sources which can be utilized for data collection are:-

* Onlines sources like **Kaggle** and **UCI** repository have datasets for loan default prediction and loan eligibility prediction. These datasets could be used for setting a benchmark for the collected datasets. These datasets represent mainly american population, so they cannot be utilized as such for the target population.
* Payment data of various individuals could be collected from various **agencies**. These include **Credit Bureau’s** payment history data , credit score data and bankruptcy data.
* Other data related to the person, such as marital status and job information could be collected using a **UI** or some other **Governmental agencies**.

## **Architecture - Flow Diagram**



**Flow Diagram - Explanation**

* **Data wrangling** - The process of cleaning and unifying messy and complex data sets for easy access and analysis.
* **EDA** - Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to.
* **Modeling** - The process of modeling means training a machine learning algorithm to predict the labels from the features, tuning it for the business need, and validating it on holdout data. The output from modeling is a trained model that can be used for inference, making predictions on new data points.
* **Model Evaluation** - After data splits into Train & Test and predicts the result.
* **Dockerizing** - An application is the process of converting an application to run within a Docker container.
* **Deployment** - 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.

## **Data Mining**

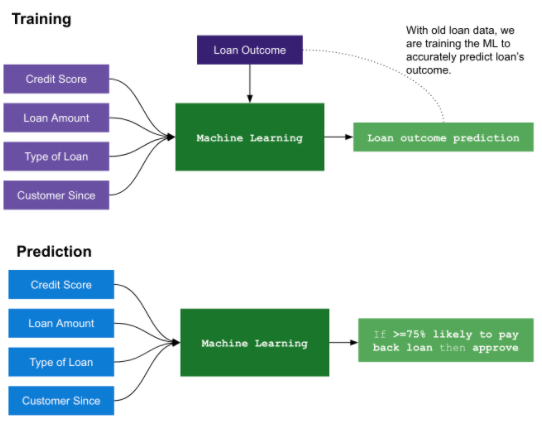
Data mining is the process of extracting useful information from large amounts of data.

Different Data Mining techniques :

* Classification
* Regression
* Clustering
* Association rule learning
* Anomaly detection

The given problem of identifying whether or not a loan should be approved is a case of **classification**. With the given data various ML algorithms like Decision Tree classifier, Support vector Machines, K nearest neighbour can be used.

For greater control of model variance ensemble approaches like Bagging (Random Forest Classifier) or Boosting (Gradient Boosting classifier , Adaboost) could be used.

****

**Simple overview of how Machine learning model will be trained and then used for the prediction of the loan acceptance/rejection outcome**

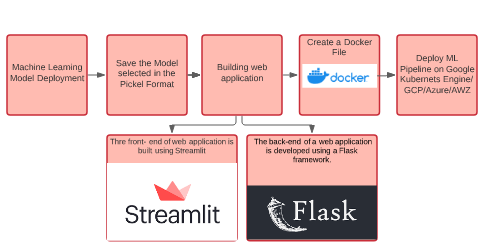
## **Technical Stack**

The technical landscape of the automated loan acceptance system is shown below:

## 

## **Deployment Strategy**

The deployment of machine learning models is the process of making models available in production where web applications, enterprise software and APIs can consume the trained model by providing new data points and generating predictions.



**Step1 — Model building**

The machine learning model which is finalized is saved as a pickle file.

**Step2 — Building Web Application**

Now that our machine learning pipeline and model are ready we will start building a web application that can connect to them and generate predictions on new data in real-time. This application will support ‘Online’ as well as ‘Batch’ predictions through a csv file upload.There are two parts of this application:

**Front-end (designed using Streamlit)** - Steamlit is an open-source Python library that makes it easy to build beautiful custom web-apps for machine learning and data science.

**Back-end (developed using Flask in Python)**-

The back-end of a web application is developed using a Flask framework.It is a framework that allows you to build web applications. A web application can be a commercial website, a blog, e-commerce system, or an application that generates predictions from data provided in real-time using trained models. Before we publish the application on Google Cloud platforms to test the web app locally. Open Anaconda Prompt and navigate to the folder where ‘app.py’ is saved on your computer and run the python file python app.py.Once executed, copy the URL into a browser ,and it should open a web application hosted on your local machine (127.0.0.1). Try entering test values to see if the predict function is working.

**Step 3 — Create a Dockerfile**

A container is a type of software that packages up an application and all its dependencies so the application runs reliably from one computing environment to another.Docker is a company that provides software (also called Docker) that allows users to build, run and manage containers while Docker’s containers are the most common.To containerize our application for deployment we need a docker image that becomes a container at runtime. A docker image is created using a Dockerfile. A Dockerfile is just a file with a set of instructions.The Dockerfile is case-sensitive and must be in the project folder with the other project files.

**Step 4 —Deploy a ML pipeline on GKE/Cloud(AWS/GCP/Azure):**

Google Kubernetes is a powerful open-source system developed by Google back in 2014, for running and managing containerized applications across a cluster of applications.

1. Sign-in to GCP console and go to Manage Resources and then click on Create New Project followed by importing the Project Code.
2. Set Environment Variable
3. Build the docker image of the application followed by authenticating to the Container Registry.
4. Create Cluster
5. Deploy Application -To deploy and manage applications on a GKE cluster, you must communicate with the Kubernetes cluster management system
6. Expose the application to the internet.
7. Check Service

## **Conclusion**

An automated system called the loan approval prediction system that will help the organization in making the decision of approving or rejecting the loan application in a quick manner.