# **Lending Club Data Analysis**

# 

ABSTRACT:

The purpose of this project is to build a data pipeline that programmatically downloads the most up-to-date data from Lending Club and train a model that can predict the interest rate which can be offered to a prospective borrower. To further validate our prediction, we shall compare the predicted interest rate for the prospective borrower with the interest rates of similar profiles from within the prospective borrower’s locality so that we can visualize and understand our predicted rate against the training set of similar profiles.

# 

# 

# **PROJECT GOALS**

## Part - 1 Requirements

* Programmatically download the data from Lending Club
* Conduct EDA to understand the data and determine the most effective features
* Pre-Process the data
* Aggregate the results in a meaningful way so that a single value is obtained
* Return a list of similar profiles from the training set based on the prospective borrower’s features
* Further filter the list of people by the locality of the prospective borrower and compare his predicted interest rate with the interest rate of all the other people in their locality

## Part - 2 Requirements

* Update the data from Lending Club’s API as and when it changes, also think of Change Data Capture strategy to download only deltas
* Implement Model Explainability and Interpretability functionality
* Scale the system using parallelization and scheduling using Dask, Luigi, Airflow, etc
* Visualize the result of the predicted rates and localities’ rates on a geo-spatial map
* Think of model governance and roll-back once the model has been retrained

# 

# 

# 

# **DATASET**

**Link to Kaggle Dataset :** <https://www.kaggle.com/wendykan/lending-club-loan-data>

There are approximately 100,000+ records and 150 features for each quarter and the data is available from 2007 until today.

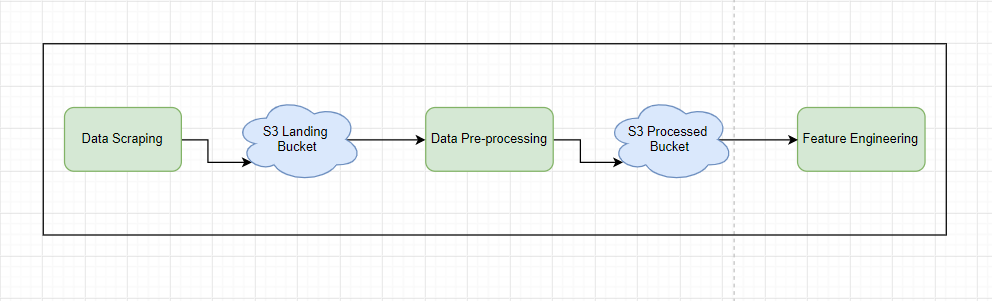
There are certain schema changes which the data underwent between 2007 and 2015, which is why we have limited our analysis to years 2016 and above.

# 

# 

# **DATA PIPELINE**

## Data Ingestion Pipeline



**STEP - 1 : Data Scraping**

The data set was programmatically downloaded from Lending Club’s official site.

Selenium was used to create a browser instance which then mimicked a user login and downloaded all the data from the site.

The zipped files are then extracted and pushed to a **Landing S3 Bucket** which serves as the source for the processing aspect of the pipeline.

**STEP - 2 : Data Pre-Processing**

The raw data obtained was first resized from the original 150 columns down to 100 columns, since all other features had a lot of empty columns.

The data types of columns that do not have any inferable type are appropriately reassigned.

There is a lot of Null data within the files and they have been handled in three ways :

* Mode Imputation : For categorical columns only.
* Mean Imputation : For Float/Integer columns only.
* Datetime Imputation : For Datetime Columns only.

# 

# 

# 

# 

# 

# Feedback from Professor

**(4th October 2019)**

**Data Lineage**

**Staging**

**Missing Values**

**Define filtering condition for Staging**

**Feedback from Professor (11th October 2019)**

**Parallelizing Ingestion**

**SparkML for Xfmations**

**S3 AWS Lambda functions**

**Maintain folder structure**

# Running Instructions

**DockerHub Commands for Assignment 1:**

Step - 1:

docker pull tiwariabhi/pgdp-a1

Step - 2:

docker run --network=host tiwariabhi/pgdp-a1 python /src/Main.py <AWS Secret> <AWS Access> <lending-club-email> <lending-club-pass>