# VISVESVARAYA TECHNOLOGICAL UNIVERSITY



JNANA SANGAMA, BELAGAVI – 590 018

**An Internship Project Report on**

***Predict The Price Of Books***

Submitted in partial fulfilment of the requirements for the VII Semester of degree of **Bachelor of Engineering in Information Science and Engineering** of Visvesvaraya Technological University, Belagavi

**by**

### Lakshmi Sarvani B Vaishnavi S M 1RN19IS402 1RN18IS115

**Under the Guidance of**

### Mr. Karthik S Chouri

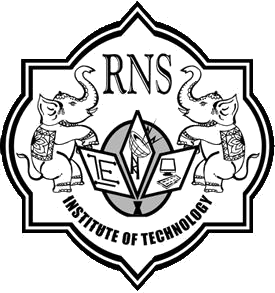
**Associate Professor Department of ISE**

**Department of Information Science and Engineering RNS Institute of Technology**

### Dr. Vishnuvaradhan Road, Rajarajeshwari Nagar post, Channasandra, Bengaluru-560098

**2021-2022**

RNS INSTITUTE OF TECHNOLOGY



### Dr. Vishnuvaradhan Road, Rajarajeshwari Nagar post, Channasandra, Bengaluru - 560098

**DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING**

**CERTIFICATE**

Certified that the Internship work entitled ***Predict The Price Of Books*** has been successfully completed by **Lakshmi Sarvani B (1RN19IS402)** and **Vaishnavi S M (1RN18IS115)** bonafide students of **RNS Institute of Technology, Bengaluru** in partial fulfilment of the requirements of 7th semester for the award of degree in **Bachelor of Engineering in Information Science and Engineering** of **Visvesvaraya Technological University, Belagavi** during academic year **2021-2022**. The internship report has been approved as it satisfies the academic requirements in respect of internship work for the said degree.

|  |  |  |
| --- | --- | --- |
| **Mr. Karthik S Chouri** | **Dr. Suresh L** | **Dr. M K Venkatesha** |
| Internship Guide | Professor and HoD | Principal |
| Associate Professor | Department of ISE | RNSIT |
| Department of ISE | RNSIT |  |
|  | **External Viva** |  |
| **Name of the Examiners** |  | **Signature with Date** |
| **1.** |  | **1. \_** |
| **2.** |  | **2. \_** |

# DECLARATION

We, **Lakshmi Sarvani B [USN: 1RN19IS402]** and **Vaishnavi S M [USN: 1RN18IS115]** students of VII Semester BE, in Information Science and Engineering, RNS Institute of Technology hereby declare that the Internship work entitled ***Predict The Price Of Books*** has been carried out by us and submitted in partial fulfilment of the requirements for the *VII Semester degree of* ***Bachelor of Engineering in Information Science and Engineering*** *of Visvesvaraya Technological University, Belagavi* during academic year 2021-2022.

Place : Bengaluru Date :

#### LAKSHMI SARVANI B

**(1RN19IS402)**

**VAISHNAVI S M**

**(1RN18IS115)**

**ABSTRACT**

As each and every sector of the market is growing, data is building up day by day, we need to keep the record of the data which can be helpful for the analytics and evaluation. Now we don’t have data in gigabyte or terabyte but in zeta byte and petabyte and this data cannot be handled with the day to day software such as Excel or Matlab. Therefore in this report we will be dealing with large data sets with the high-level programming language ’Python’.

The main goal of this project is to aggregate and analyze the data collected from the different data sets. This project mainly focuses on the usage of the python programming language in the field of data analysis and outcome prediction. This language has not only it’s application in the field of just analyzing the data but also for the prediction of the upcoming scenarios.

The purpose of using this specific language is due to its versatility, vast libraries (Pandas, Numpy, Matplotlib, etc.), speed limitations, and ease of learning. We will be analyzing large data sets in this project which can not be easily analyzed in other tools as compared to python. Python does not have it’s limitation to only data analytics but also in many other fields such as Artificial intelligence, Machine learning, and many more.

i

# ACKNOWLEDGMENT

At the very onset I would like to place our gratefulness to all those people who helped me in making the Internship a successful one.

Coming up, this internship to be a success was not easy. Apart from the sheer effort, the enlightenment of the very experienced teachers also plays a paramount role because it is they who guided me in the right direction.

First of all, I would like to thank the **Management of RNS Institute of Technology** for providing such a healthy environment for the successful completion of internship work.

In this regard, I express sincere gratitude to our beloved Principal **Dr. M K Venkatesha,** for providing us all the facilities.

We are extremely grateful to our own and beloved Professor and Head of Department of Information science and Engineering, **Dr. Suresh L**, for having accepted to patronize me in the right direction with all her wisdom.

We place our heartfelt thanks to **Mr. Karthik S Chouri** Assistant Professor, Department of Information Science and Engineering for having guided internship and all the staff members of the department of Information Science and Engineering for helping at all times.

I thank **Mr. Satyendra Nath Data Scientist at LocalHost Tecnology**, for providing the opportunity to be a part of the Internship program and having guided me to complete the same successfully.

I also thank our internship coordinator **Dr. R Rajkumar,** Associate Professor, Department of Information Science and Engineering. I would thank my friends for having supported me with all their strength and might. Last but not the least, I thank my parents for supporting and encouraging me throughout. I have made an honest effort in this assignment.

ii

# TABLE OF CONTENTS

|  |  |
| --- | --- |
| **Abstract** | **i** |
| **Acknowledgment** | **ii** |
| **Contents** | **iii** |
| **List of figures** | **v** |
| **List of Abbreviations** | **vi** |
| **1. Introduction** | **1** |
| 1.1 Data Science | 1 |
| 1.2 Existing System | 1 |
| 1.3 Proposed System | 2 |
| **2. Analysis** | **3** |
| 2.1 Introduction | 3 |
| 2.1.1 Flutter | 3 |
| 2.1.2 Flutter REST API | 4 |
| 2.2 Software Requirement Specification | 5 |
| **3. System Design** | **6** |
| 3.1 Flutter Framework | 6 |
| 3.1.1 Basic Widgets | 7 |
| 3.1.2 What is Material Design? | 8 |
| 3.2 Flutter Architecture | 10 |
| iii |  |

|  |  |
| --- | --- |
| **4. Detailed Design** | **11** |
| 4.1 Process Flow | 11 |
| 4.2 Screen Design | 11 |
| **5. Implementation Details** | **17** |
| 5.1 Implementation | 17 |
| 5.2 Implementing REST API in Flutter | 22 |
| **6. Testing** | **26** |
| 6.1 Common Flutter Errors | 26 |
| 6.2 Handling Errors in REST API | 28 |
| **7. Results** | **31** |
| **8. Conclusion and Future Work** | **37** |
| 8.1 Conclusion | 37 |
| 8.2 Future Work | 37 |
| **9. References** | **38** |
| = iv |  |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figures** | **Descriptions** | **Page** |
| Figure. 2.2 | XGBoost Algorithm | 12 |
| Figure. 3.1 |  | 12 |
| Figure. 3.2 |  | 07 |
| Figure. 3.3 |  | 10 |
| Figure. 3.4 |  | 10 |
| Figure. 4.1 |  | 11 |
| Figure. 4.2 |  | 12 |
| Figure. 4.3 |  | 12 |
| Figure. 4.4 |  | 13 |
| Figure. 4.5 |  | 13 |
| Figure. 4.6 |  | 14 |
| Figure. 4.7 |  | 14 |
| Figure. 4.8 |  | 15 |
| Figure. 4.9 |  | 16 |
| Figure. 4.10 |  | 16 |
| Figure. 5.1 |  | 22 |
| Figure. 6.1 |  | 26 |
| Figure. 6.2 |  | 27 |
| Figure. 6.3 |  | 27 |
| Figure. 6.4 |  | 28 |
| Figure. 7.1 |  | 31 |
| Figure. 7.2 |  | 31 |
| Figure. 7.3 |  | 31 |
| Figure. 7.4 |  | 32 |
| Figure. 7.5 |  | 32 |
| Figure. 7.6 |  | 32 |
| Figure. 7.7 |  | 32 |
| Figure. 7.8 |  | 33 |
| Figure. 7.9 |  | 33 |
| Figure. 7.10 |  | 33 |
| Figure. 7.11 |  | 33 |
| Figure. 7.12 |  | 34 |
| Figure. 7.13 |  | 34 |
| Figure. 7.14 |  | 35 |
| Figure. 7.15 |  | 36 |

v

# LIST OF ABBREVATIONS

|  |  |  |
| --- | --- | --- |
| SQL | - | Structured Query Language |
| SAS | - | Static Analysis System |
| API | - | Application programming interface |
|  | - |  |
|  | - |  |
|  | - |  |
|  | - |  |
|  | - |  |
|  | - |  |
|  | - |  |
|  | - |  |
|  | - |  |
|  | - |  |

vi

# INTRODUCTION

## Data Science

Data science is the field of data analytics and data visualization in which raw data or the unstructured data is cleaned and made ready for the anlysis purpose. Data scientists use this data to get the required information for the future purpose. ”Data science uses many processes and methods on the big data, the data may be structured or unstructured”. Data frames available on the internet is the raw data we get. It may be either in unstructured or semi structured format. This data is further filtered, cleaned and then number of required task are performed for the analysis with the use of the high programming language. This data is further analyzed and then presented for our better understanding and evaluation.

One must be clear that data science is not about making complicated models or making awesome visualization neither it is about writing code but about using the data to create an impact for your company, for this impact we need tools like complicated data models and data visualization. There are many tools used to handle th**e big data available to us.** ”Data scientists use programming tools such as Python, R, SAS, Java, Perl, and C/C++ to extract knowledge fromprepared data”.

## Data scientists use many algorithms and mathematical models on the data. Following are the stages and their cycle performed on the unstructured data.

## • Identifying the problem.

## • Identify available data sources

## • Identify available data sources

## • Identify if additional data sources are needed.

## • Statistical analysis

## • Implementation, development

## • Communicate results

## • Maintenance

## Data science finds its application in many fields. With the assistance of data science it is easy to get the search query on search engines in plenty of time. A role of the data scientist is to have a deep understanding of the data as well as a good command on the programming language, he should also know how to work with the raw data extracted from the data source. Many programming languages are used to analyze and evaluate the data such as Python, Java, MATLAB, Scala, Julia, R., SQL and TensorFlow. Among which python is the most user friendly and vastly used programming language in the field of data science. This life cycle is applied in each and every field, in this project we will be considering all this seven stages of data science to analyze the data. The process will be starting from data collection, data preparation, data modeling and finally data evaluation.

## Existing System

## Proposed System

## The proposed system uses XGBoost for regression. Extreme Gradient Boosting (XGBoost) is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm. Shortly after its development and initial release, XGBoost became the go-to method and often the key component in winning solutions for a range of problems in machine learning competitions. Regression predictive modeling problems involve predicting a numerical value such as a dollar amount or a height. XGBoost can be used directly for regression predictive modeling.

## Gradient boosting refers to a class of ensemble machine learning algorithms that can be used for classification or regression predictive modeling problems. Ensembles are constructed from decision tree models. Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models. This is a type of ensemble machine learning model referred to as boosting. Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm. This gives the technique its name, “gradient boosting,” as the loss gradient is minimized as the model is fit, much like a neural network. The two main reasons to use XGBoost are execution speed and model performance.

# ANALYSIS

## Introduction

### ”Python is an interpreted, object-oriented, high-level programming language with dynamic semantics”. This language consist of mainly data structures which make it very easy for the data scientists to analyse the data very effectively. It does not only help in forecasting and analysis it also helps in connecting the two different languages.Two best features of this programming language is that it does not have any compilation step as compared to the other programming language in which compilation is done before the program is being executed and other one is the reuse of the code, it consist of modules and packages due to which we can use the previously written code any where in between the program whenever is required. There are multiple languages for example R., Java, SQL, Julia, Scala, MATLAB available in market which can be used to analyze and evaluate the data, but due to some outstanding features python is the most famous language used in the field of data science. Python is mostly used and easy among all other programming languages.

## XGBoost Regression

## Extreme Gradient Boosting, or XGBoost for short, is an efficient open-source implementation of the gradient boosting algorithm. As such, XGBoost is an algorithm, an open-source project, and a Python library. It was initially developed by Tianqi Chen and was described by Chen and Carlos Guestrin in their 2016 paper titled “XGBoost: A Scalable Tree Boosting System.” It is designed to be both computationally efficient (e.g. fast to execute) and highly effective, perhaps more effective than other open-source implementations. The two main reasons to use XGBoost are execution speed and model performance. XGBoost dominates structured or tabular datasets on classification and regression predictive modeling problems. The evidence is that it is the go-to algorithm for competition winners on the Kaggle competitive data science platform.

## 

***Fig 2.2 XGBoost Algorithm***

## XGBoost Regression API

## XGBoost can be installed as a standalone library and an XGBoost model can be developed using the scikit-learn API. The first step is to install the XGBoost library if it is not already installed. This can be achieved using the pip python package manager on most platforms; for example:

## sudo pip install xgboost

You can then confirm that the XGBoost library was installed correctly and can be used by running the following script.

import xgboost

print(xgboost.\_\_version\_\_)

Running the script will print your version of the XGBoost library you have installed.

Your version should be the same or higher. If not, you must upgrade your version of the XGBoost library.

It is possible that you may have problems with the latest version of the library. It is not your fault.

Sometimes, the most recent version of the library imposes additional requirements or may be less stable.

If you do have errors when trying to run the above script, I recommend downgrading to version 1.0.1 (or lower). This can be achieved by specifying the version to install to the pip command, as follows:

sudo pip install xgboost==1.0.1

The XGBoost library has its own custom API, although we will use the method via the scikit-learn wrapper classes: XGBRegressor and XGBClassifier. This will allow us to use the full suite of tools from the scikit-learn machine learning library to prepare data and evaluate models.

An XGBoost regression model can be defined by creating an instance of the XGBRegressor class; for example:

...

# create an xgboost regression model

model = XGBRegressor()

You can specify hyperparameter values to the class constructor to configure the model. Perhaps the most commonly configured hyperparameters are the following:

* n\_estimators: The number of trees in the ensemble, often increased until no further improvements are seen.
* max\_depth: The maximum depth of each tree, often values are between 1 and 10.
* eta: The learning rate used to weight each model, often set to small values such as 0.3, 0.1, 0.01, or smaller.
* subsample: The number of samples (rows) used in each tree, set to a value between 0 and 1, often 1.0 to use all samples.
* colsample\_bytree: Number of features (columns) used in each tree, set to a value between 0 and 1, often 1.0 to use all features.

## Software requirement specification

The best thing about using Flutter for creating cross-platform native mobile apps is the fact that you can build those on almost any OS.

Here are some System Requirements for Android Studio which is needed for running an Android simulator.

#### Windows:

* Microsoft® Windows® 7/8/10 (64-bit)
* 4 GB RAM minimum, 8 GB RAM recommended
* 2 GB of available disk space minimum,
* 4 GB Recommended (500 MB for IDE + 1.5 GB for Android SDK and emulator system image)
* 1280 x 800 minimum screen resolution

To install and run Jupyter, your development environment must meet these minimum requirements:

* **Operating Systems**: Windows 7 SP1 or later (64-bit), x86-64 based.
* **Disk Space**: 1.64 GB (does not include disk space for IDE/tools).
* With PySpark (Team Studio version 6.2 and later)

Memory and disk space required per user: 1GB RAM + 1GB of disk + .5 CPU core.

Server overhead: 2-4GB or 10% system overhead (whatever is larger), .5 CPU cores, 10GB disk space.

Port requirements: Port 8000 plus 5 unique, random ports per notebook.

* Without PySpark (Team Studio version 6.0 or 6.1)

Memory and disk space required per user: 512MB RAM + 1GB of disk + .5 CPU core.

Server overhead: 2-4GB or 10% system overhead (whatever is larger), .5 CPU cores, 10GB disk space.

Port requirements: Port 8000.SYSTEM DESIGN

# DETAILED DESIGN

## Process Flow

## 

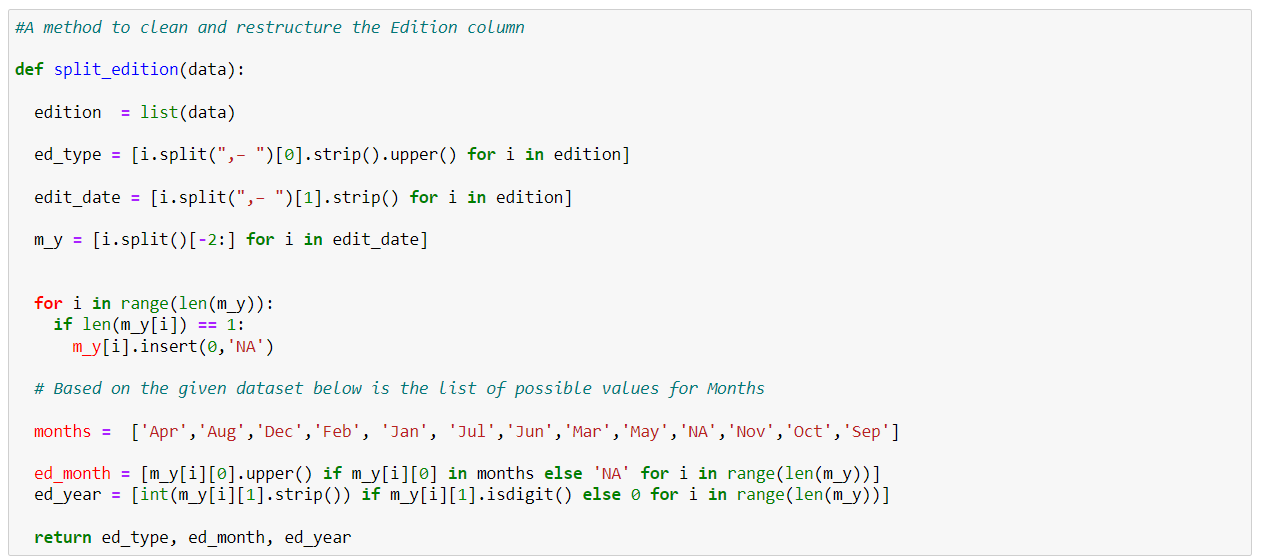
## 

# IMPLEMENTATION DETAILS

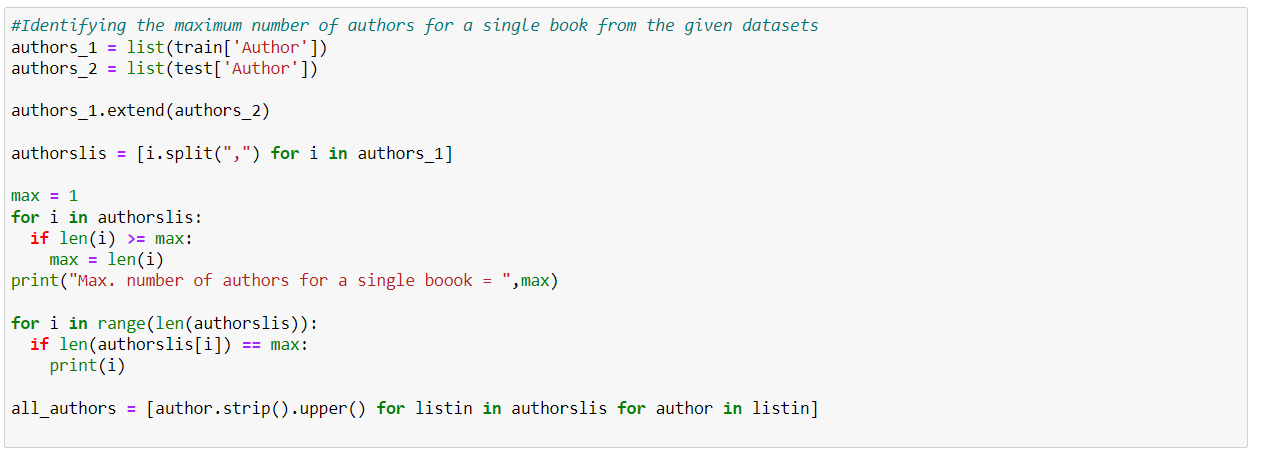
## Implementation

## 

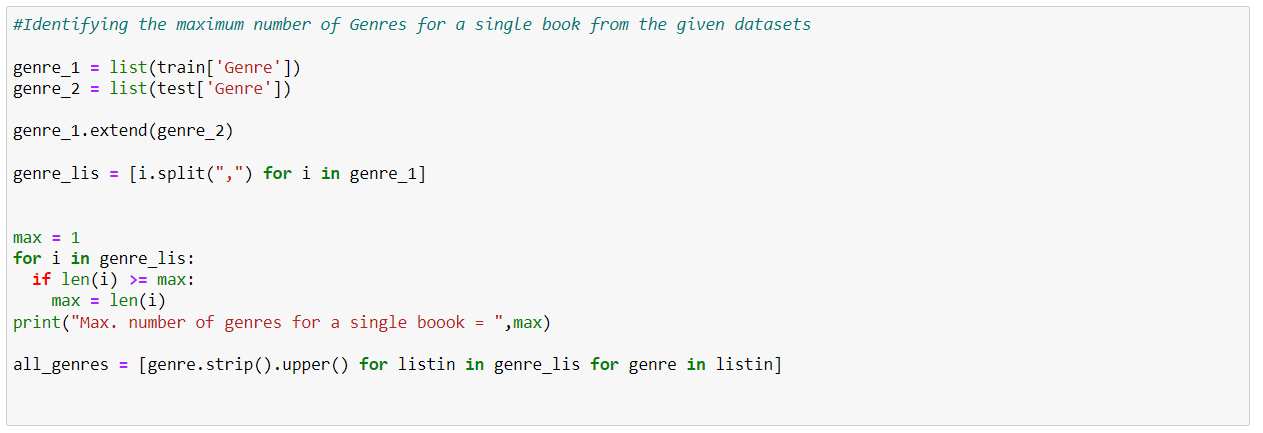
The snippet is used to import the training and test data sets.

****

The snippet is used to clean and restructure the edition column.



The snippet is used to identify the maximum number of authors for a single book from given data set.



The snippet is used to to identify the maximum number of genres for a single book from given data set.



The snippet is implemented to split the genre column into 7 new columns.



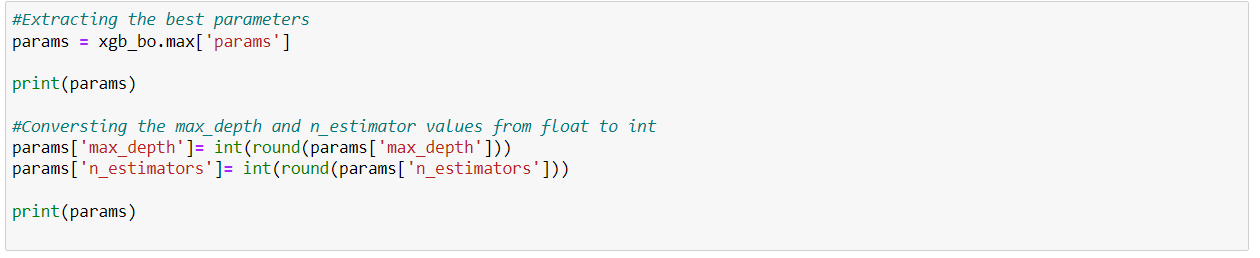
The snippet is used to clean and restructure the datasets.



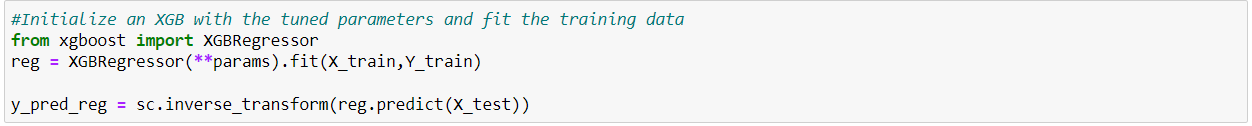
The snippet is used to install xgboost.



The snippet is used to install bayesian optimization.



The snippet is used to extract the best parameters.



The snippet to initialize an XGB with the tuned parameters and fit the raining data.

# TESTING

Testing is the process of evaluating and verifying that a software product or application does what it is supposed to do.

# RESULTS

****

The predicted prices of the books.

# CONCLUSION AND FUTURE WORK

## Conclusion

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable.

## Future Work

## We can further increase the efficiency of the algorithm.

## Improve the accuracy of prediction.

## Try to decrease the execution time.

# REFERENCES

* <https://machinehack.com/hackathon/predict_the_price_of_books/data>
* <https://medium.com/analytics-vidhya/books-price-prediction-via-python-31dc358ad8d8>
* <https://machinelearningmastery.com/xgboost-for-regression>
* <https://www.geeksforgeeks.org/xgboost-for-regression>