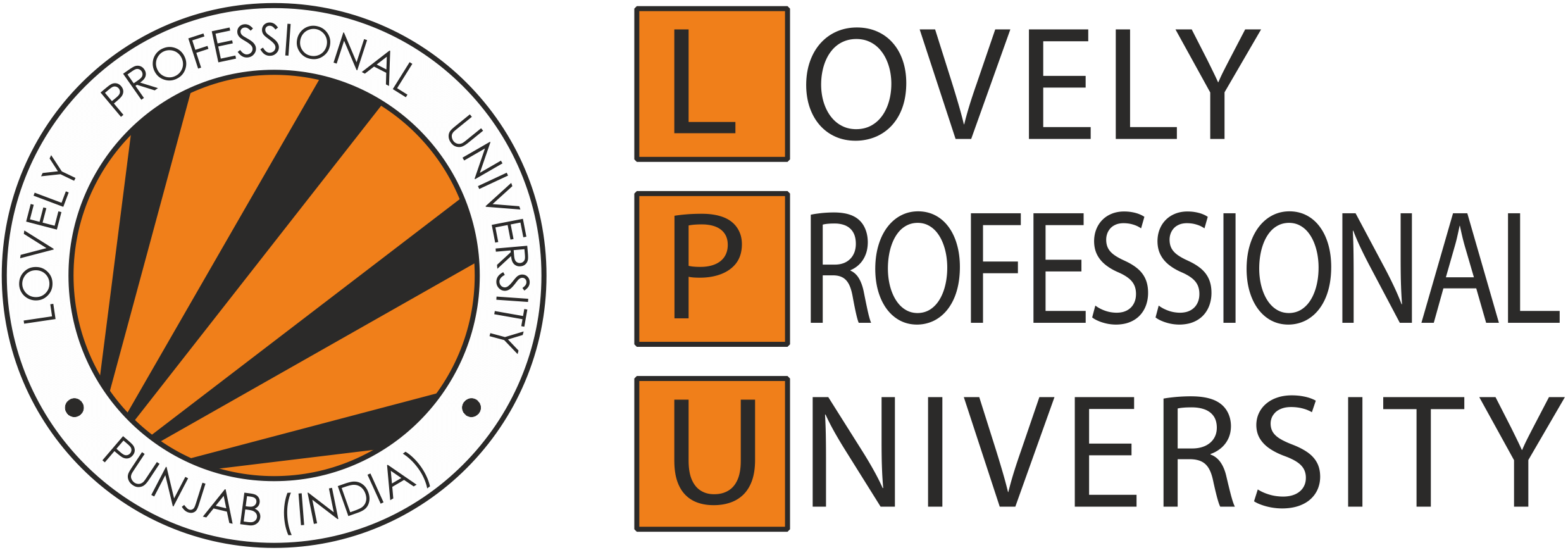
**PROJECT REPORT**

**Submitted to: - LOVELY PROFESSIONAL UNIVERSITY PHAGWARA,**



**House Price prediction**

**SUBMITTED BY** -:

**Name of Student**: Armaan Srivastava

**Registration Number**: 11910720

**Signature of Student**: *Armaan Srivastava*

**TABLE OF CONTENTS**

**Chapter Number Contents**

INTRODUCTION

* 1. AIM and IMPORTANCE
     1. Aim
     2. Need and Motivation

1. DATASET
   1. Steps in Preparing Data for Model
      1. Data Exploration
      2. Data Visualization
      3. Data Selection
      4. Data Transformation
2. LANGUAGE AND MODELS USED
   1. Python
      1. Jupyter Notebook
      2. NumPy
      3. Pandas
      4. Seaborn
      5. Matplotlib
3. **Models Used**
   * 1. Random Forest Regressor
     2. Deep learning

### LIST OF FIGURES

|  |  |  |  |
| --- | --- | --- | --- |
| FIGURE | TITLE |  |  |
| 4.1 | Bar Plots of various Parameters |  |  |
| 4.2 | Correlation Heatmap |  |  |
| 4.3 | Train test split |  |  |
| 4.4 | Random forest regressor |  |  |
| 4.5 | Keras epoch |  |  |
| 4.6 | Model summary |  |  |

### ABBREVIATIONS

SqM Square meter

SqFt Square Feet

LR Linear Regression

rmse Root Mean Square Error Log Logarithmic

# Abstract

House price forecasting is an important topic of real estate. The literature attempts to derive useful knowledge from historical data of property markets. Machine learning techniques are applied to analyze historical property transactions in India to discover useful models for house buyers and sellers. Revealed is the high discrepancy between house prices in the most expensive and most affordable suburbs in the city of India. Moreover, experiments demonstrate that the Multiple Linear Regression that is based on mean squared error and deep learning measurement is a competitive approach.

# INTRODUCTION

AIM and IMPORTANCE

**Aim**

**These are the Parameters on which we will evaluate ourselves-**

* Create an effective price prediction model
* Validate the model’s prediction accuracy
* Identify the important home price attributes which feed the model’s predictive power.

## Need and Motivation

Having lived in India for so many years if there is one thing that I had been taking for granted, it’s that housing and rental prices continue to rise. Since the housing crisis of 2008, housing prices have recovered remarkably well, especially in major housing markets. However, post-covid-19, I was surprised to read that the housing prices had fallen the most in the last 4 years. In fact, median resale prices for condos and coops fell 6.3%, marking the first time there was a decline since Q3 of 2019. The decline has been partly attributed to political uncertainty domestically and abroad and the 2019 election. So, to maintain the transparency among customers and also the comparison can be made easy through this model. If customer finds the price of house at some given website higher than the price predicted by the model, so he can reject that house.

# DATASET

Here we have web scrapped the Data from 99acres.com website which is one of the leading real estate websites operating in INDIA.

## Dataset looks as follows-

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated

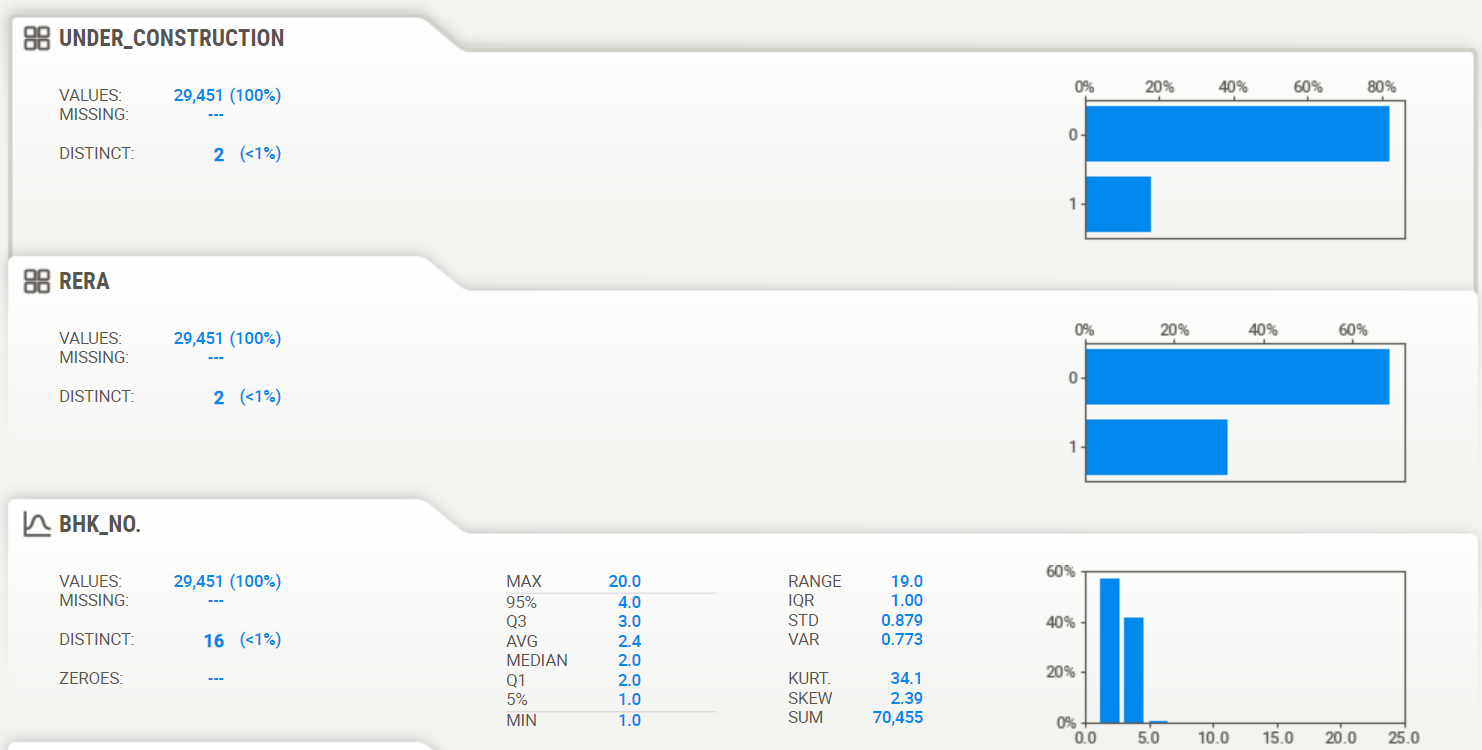
**Data Exploration**

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python**.** Before it can conduct analysis on data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working.

We divided the data 9:1 for Training and Testing purpose respectively.

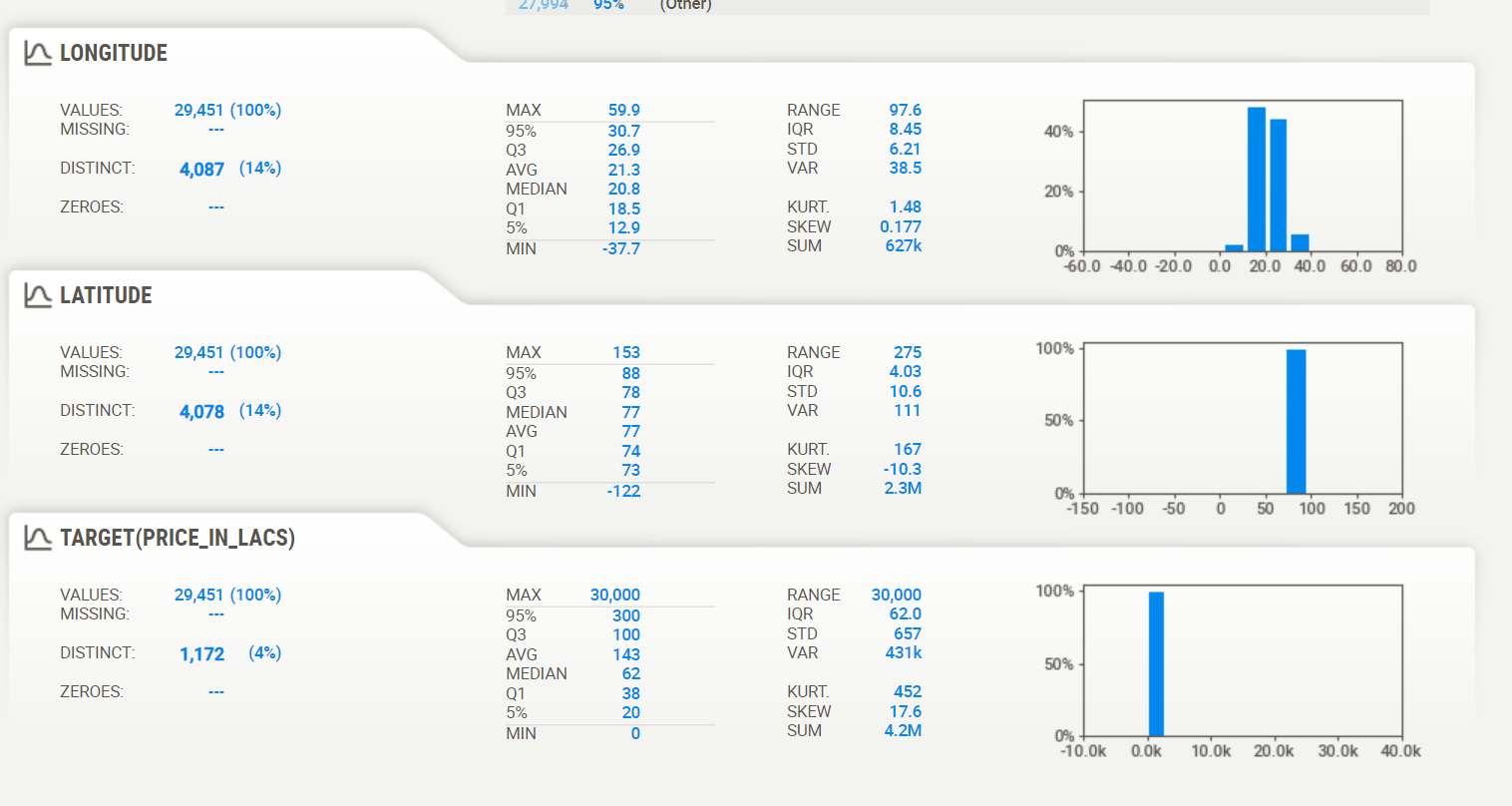
## Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies are essential to analyses massive amounts of information and make data-driven decisions.



Graphical user interface, application

Description automatically generated



**Data Selection**

Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Data selection precedes the actual practice of data collection. This definition distinguishes data selection from selective data reporting (selectively excluding data that is not supportive of a research hypothesis) and interactive/active data selection (using collected data for monitoring activities/events, or conducting secondary data analyses). The process of selecting suitable data for a research project can impact data integrity.

The primary objective of data selection is the determination of appropriate data type, source, and instrument(s) that allow investigators to adequately answer research questions. This determination is often discipline-specific and is primarily driven by the nature of the investigation, existing literature, and accessibility to necessary data sources.

**Correlation Heatmap**

Chart, treemap chart

Description automatically generated

**Data Transformation**

The log transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics.

It is hard to discern a pattern in the upper panel whereas the strong relationship is shown clearly in the lower panel. The comparison of the means of log-transformed data is actually a comparison of geometric means. This occurs because, as shown below, the anti-log of the arithmetic mean of log-transformed values is the geometric mean.

**Python**

Python is widely used in scientific and numeric computing:

* SciPy is a collection of packages for mathematics, science, and engineering.
* Pandas is a data analysis and modelling library.
* IPython is a powerful interactive shell that features easy editing and recording of a work session, and supports visualizations and parallel computing.
* The Software Carpentry Course teaches basic skills for scientific computing, running

bootcamps and providing open-access teaching materials.

**Libraries Used for this Project include –**

* **Pandas**
* Pandas is an open-source, BSD-licensed Python library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.
* Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.
* **NumPy**
* NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays.
* Using NumPy, mathematical and logical operations on arrays can be performed.
* **Matplotlib**
* Matplotlib is one of the most popular Python packages used for data visualization. It is a cross-platform library for making 2D plots from data in arrays.
* It provides an object-oriented API that helps in embedding plots in applications using Python GUI toolkits such as PyQt, WxPythonotTkinter.
* **Seaborn**
* Seaborn is an open source, BSD-licensed Python library providing high level API for visualizing the data using Python programming language
* **Scikit Learn**
* Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python.
* It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

**Model**

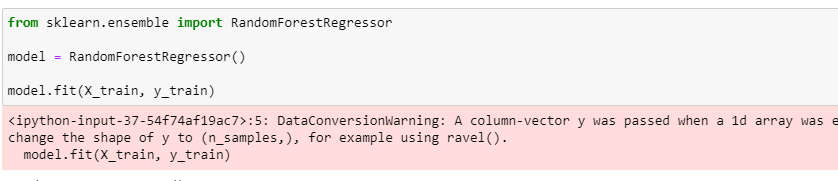
**Random Forest Regressor**

* A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging.
* Bagging, in the Random Forest method, involves training each decision tree on a different data sample where sampling is done with replacement.
* The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision tree

**Train Test split**



**Model**

****

**Prediction**

**Graphical user interface, text, application

Description automatically generated**

**KERAS**

Keras runs on top of open source machine libraries like TensorFlow. Theano is a python library used for fast numerical computation tasks. TensorFlow is the most famous symbolic math library used for creating neural networks and deep learning models. TensorFlow is very flexible and the primary benefit is distributed computing. CNTK is deep learning framework developed by Microsoft. It uses libraries such as Python, C#, C++ or standalone machine learning toolkits. Theano and TensorFlow are very powerful libraries but difficult to understand for creating neural networks.

Keras is based on minimal structure that provides a clean and easy way to create deep learning models based on TensorFlow or Theano. Keras is designed to quickly define deep learning models. Well, Keras is an optimal choice for deep learning applications.

Keras is highly powerful and dynamic framework and comes up with the following advantages −

* Larger community support.
* Easy to test.
* Keras neural networks are written in Python which makes things simpler.
* Keras supports both convolution and recurrent networks.
* Deep learning models are discrete components, so that, you can combine into many ways.

**Train Test split**

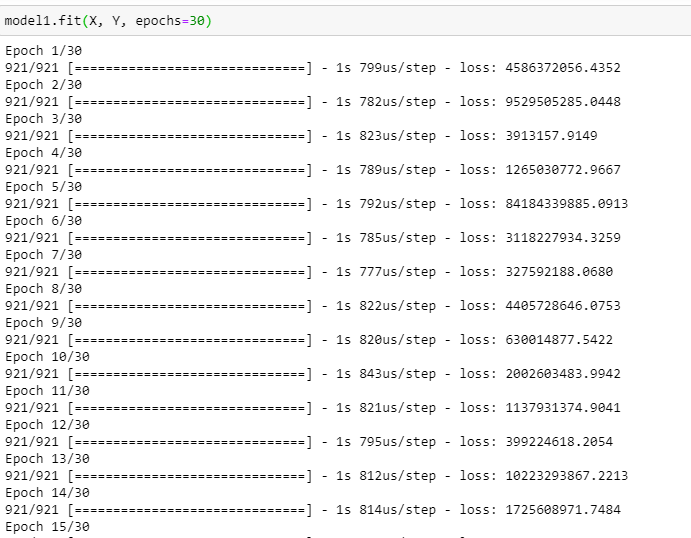


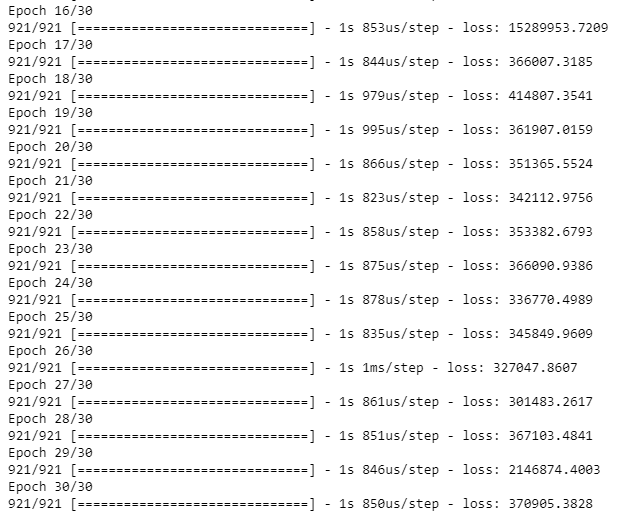
**Model**

**Graphical user interface, text, application

Description automatically generated**

**Epochs**

****

****

**Model Summary**

**Table

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**