National College of Ireland

Configuration Manual

MSc Research Project
MSc AI

Shubham Kawade
Student ID: X23354658

School of Computing
National College of Ireland

Supervisor: Prof. Sheresh Zahoor

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Shubham Pandurang Kawade

Student ID: x23354658

Programme: MSc. Artificial Intelligence **Year:** 2024-2025

Module: MSc.Research Project

Lecturer: Prof Sheresh Zahoor

Submission Due Date: 01/09/2025

Project Title: Predicting Box Office Success with Advanced Regression Techniques: Using Data to Merge Art and Analytics

Word Count: **Page Count:** 14

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Shubham kawade

Date: 01/09/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Forename Surname
Student ID:

1 Introduction

This Configuration Manual provides detailed instructions to replicate the computational environment and workflow for the MSc research project, *Predicting Box Office Success with Advanced Regression Techniques: Using Data to Merge Art and Analytics*. The project predicts box office revenue and IMDb ratings using the Kaggle IMDb Movie Dataset (1,000 films, CC-BY license) with regression (Linear, Ridge, Lasso) and tree-based models (Random Forest, Gradient Boosting, XGBoost). The manual covers system requirements, environment setup, data acquisition and preprocessing, model implementation, output generation, and troubleshooting, ensuring reproducibility for researchers and practitioners. It aligns with the thesis's Methodology (Section 3), Design Specification (Section 4), and Implementation (Section 5) chapters, using Python 3.8+ and libraries like scikit-learn and xgboost.

2 System Requirements

2.1 Hardware Requirements

- **Processor**: Intel i5/i7 or equivalent (used: Intel i7).
- **Memory**: Minimum 8GB RAM, recommended 16GB (used: 16GB).
- **Storage**: 5GB free disk space for dataset, code, and outputs.
- **Operating System**: Windows 10/11, macOS, or Linux (used: Windows/Linux).

2.2 Software Requirements

- **Python**: Version 3.8 or higher.
- **Libraries**:
  - ✓ pandas (1.5.3): Data manipulation.
  - ✓ numpy (1.24.3): Numerical computations.
  - ✓ scikit-learn (1.2.2): Regression and tree-based models, preprocessing, evaluation.
  - ✓ xgboost (2.0.3): XGBoost model implementation.
  - ✓ seaborn (0.12.2), matplotlib (3.7.1): Visualizations.
  - ✓ scipy (1.10.1): Statistical analysis.
- **Package Manager**: pip or conda for dependency installation.
- **IDE**: Jupyter Notebook, VS Code, or PyCharm (used: Jupyter Notebook).
- **Dataset**: Kaggle IMDb Movie Dataset (imdb_movie_dataset.csv), available at [Kaggle URL] under CC-BY license.

3 Environment Setup

3.1 Installing Python

1. Download Python 3.8+ from python.org.
2. Install, ensuring pip is included and added to PATH.
3. Verify installation: python --version (should return 3.8 or higher).

3.2 Installing Dependencies

1. Install required libraries using pip:

   pip install pandas==1.5.3 numpy==1.24.3 scikit-learn==1.2.2 xgboost==2.0.3 seaborn==0.12.2 matplotlib==3.7.1 scipy==1.10.1

2. Verify installations: pip list (should list all packages with correct versions).

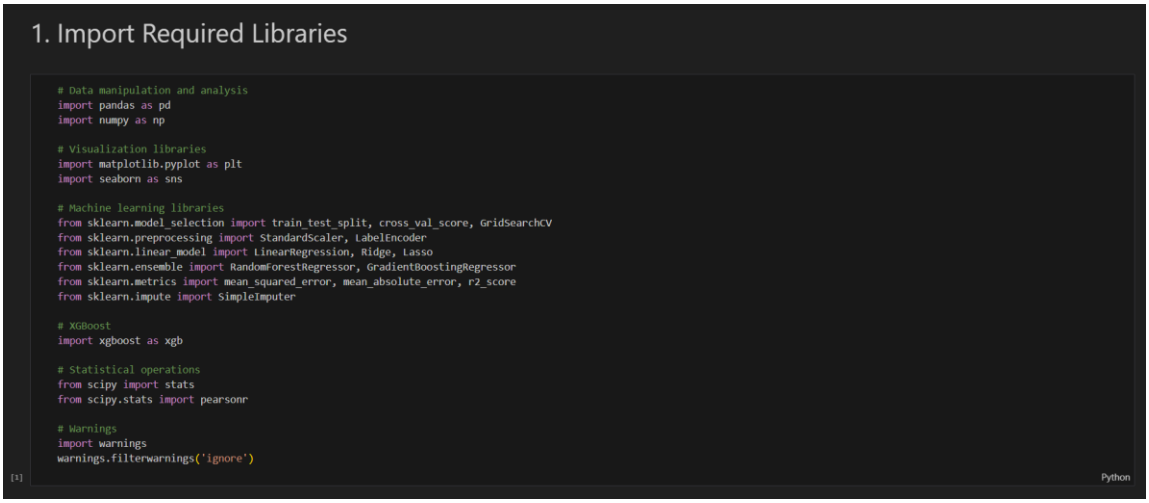

Figure 1: Importing Necessary Libraries

4 Data Acquisition and Preprocessing

4.1 Acquiring the Dataset

1. Download the Kaggle IMDb Movie Dataset from [Kaggle URL] (CC-BY license).
2. Save as imdb_movie_dataset.csv in the project directory (e.g., thesis_data/).
3. Verify file integrity: 1,000 rows, 12 columns (Rank, Title, Genre, Description, Director, Actors, Year, Runtime, Rating, Votes, Revenue, Metascore).

```python
# Load the dataset
imdb = pd.read_csv('imdb_movie_dataset.csv')

# Display the first few rows of the dataset
print("Dataset loaded successfully!")
print(imdb.head())
```

```
Dataset loaded successfully!
   Rank                    Title                     Genre  \
0     1  Guardians of the Galaxy   Action,Adventure,Sci-Fi
1     2               Prometheus  Adventure,Mystery,Sci-Fi
2     3                    Split           Horror,Thriller
3     4                     Sing   Animation,Comedy,Family
4     5            Suicide Squad  Action,Adventure,Fantasy


0                                                A group of intergalactic criminals are forced to work together to stop a fanatical war
1                                                Following clues to the origin of mankind, a team finds a structure on a distant moon, l
2                                 Three girls are kidnapped by a man with a diagnosed 23 distinct personalities. They must try to escape before the app
3  In a city of humanoid animals, a hustling theater impresario's attempt to save his theater with a singing competition becomes grander than he anticipates even as its finalists' find
4                        A secret government agency recruits some of the most dangerous incarcerated super-villains to form a defensive task force. Their first mis

             Director  \
0          James Gunn
1        Ridley Scott
2  M. Night Shyamalan
3 Christophe Lourdelet
4          David Ayer

                                              Actors  \
0                Chris Pratt, Vin Diesel, Bradley Cooper, Zoe Saldana
1   Noomi Rapace, Logan Marshall-Green, Michael Fassbender, Charlize Theron
...
1  2012          124    7.0  485820          126.46      65.0
2  2016          117    7.3  157606          138.12      62.0
3  2016          108    7.2   60545          270.32      59.0
4  2016          123    6.2  393727          325.02      40.0
```

Figure 2: Data Loading and exploration

4.2 Preprocessing Steps

```python
print("Dataset Shape:", imdb.shape)
```

```
Dataset Shape: (1000, 12)
```

```python
print("\nDataset Info:")
print(imdb.info())
```

```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Rank                1000 non-null   int64
 1   Title               1000 non-null   object
 2   Genre               1000 non-null   object
 3   Description         1000 non-null   object
 4   Director            1000 non-null   object
 5   Actors              1000 non-null   object
 6   Year                1000 non-null   int64
 7   Runtime (Minutes)   1000 non-null   int64
 8   Rating              1000 non-null   float64
 9   Votes               1000 non-null   int64
 10  Revenue (Millions)  872 non-null    float64
 11  Metascore           936 non-null    float64
dtypes: float64(3), int64(4), object(5)
memory usage: 93.9+ KB
None
```

Figure 3: Data information

```python
# Check for missing values
missing_values = imdb.isnull().sum()
missing_percentage = (missing_values / len(imdb)) * 100

missing_df = pd.DataFrame({
    'Missing Count': missing_values,
    'Missing Percentage': missing_percentage
})

print("Missing Values Analysis:")
print(missing_df[missing_df['Missing Count'] > 0])
```

```
Missing Values Analysis:
                    Missing Count  Missing Percentage
Revenue (Millions)            128                12.8
Metascore                      64                 6.4
```

```python
# Basic statistics
print("\nBasic Statistics:")
imdb.describe()
```

Figure 4: Missing Values and Summary statistics

Figure 5: Data Preprocessing and Feature Engineering

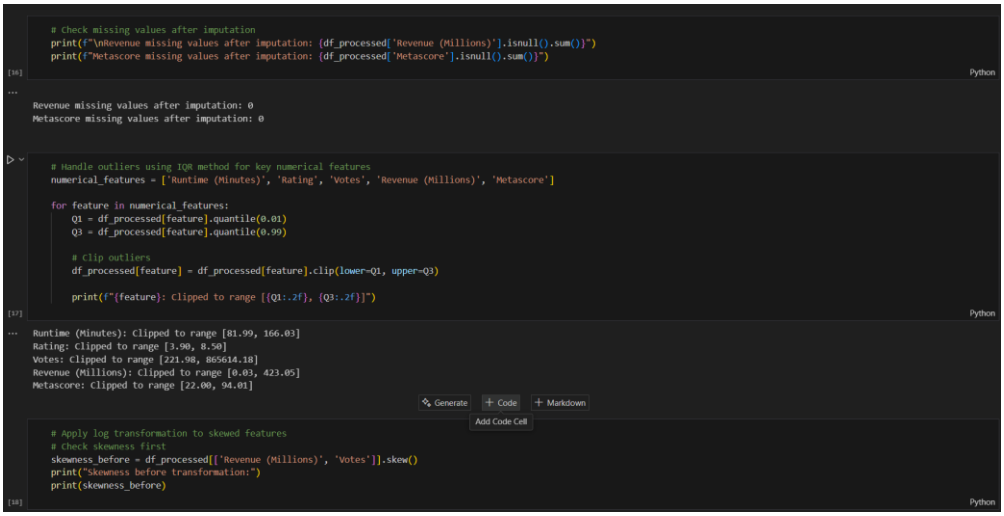

Figure 6: Handling Missing Values and Outliers

Figure 7: Apply Log Transformation

Figure 8: Generic Feature Engineering

Figure 9: Additional Feature Created

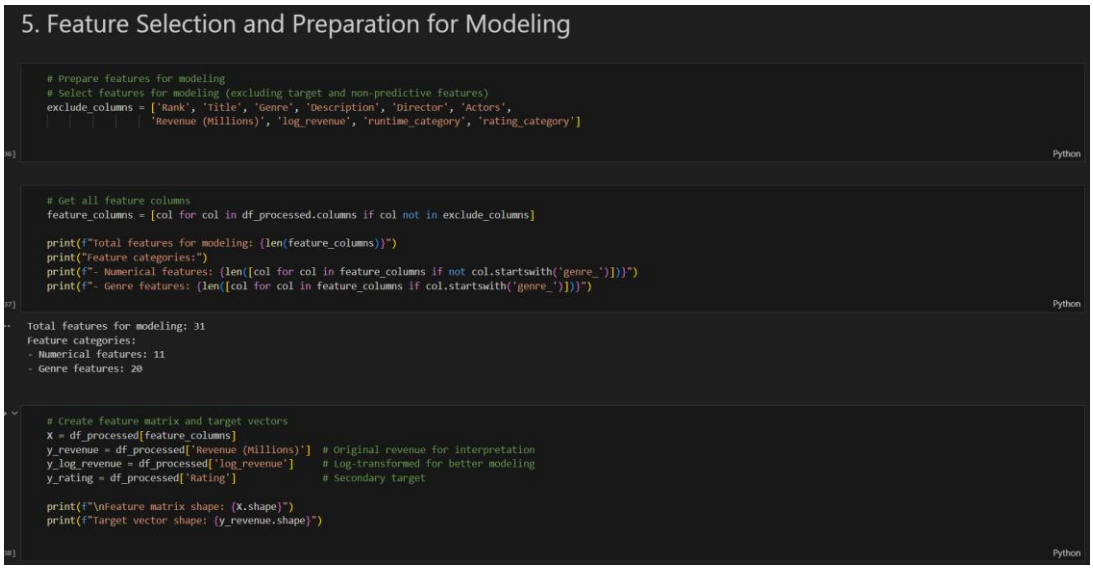

Figure 10: Feature Selection and preparing for modelling

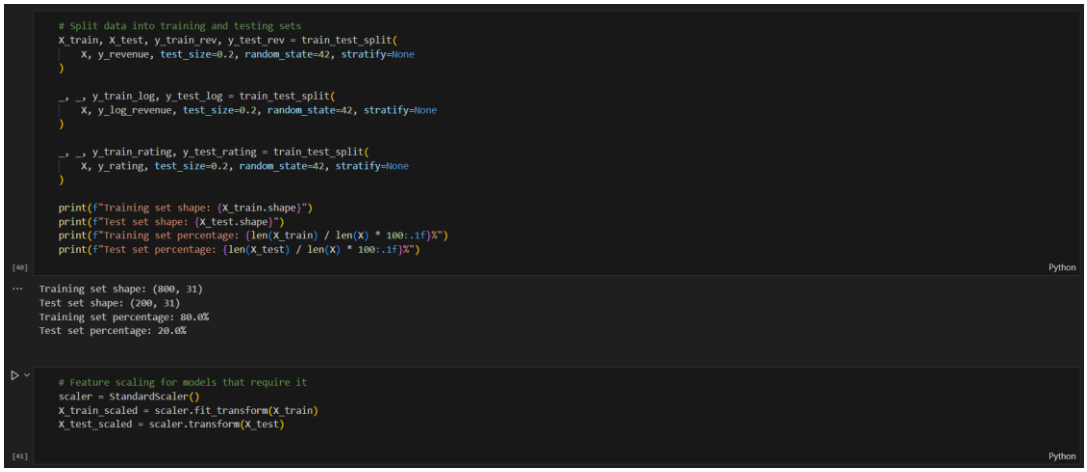

Figure 11: Splitting Train Test

5 Model Implementation

5.1 Feature Matrix and Target Vectors

1. **Revenue Prediction**:
    - Feature matrix: 31 columns (11 numeric, 20 genre).
    - Target: y_revenue = df_processed['Revenue (Millions)'] or y_log_revenue = df_processed['log_revenue'].

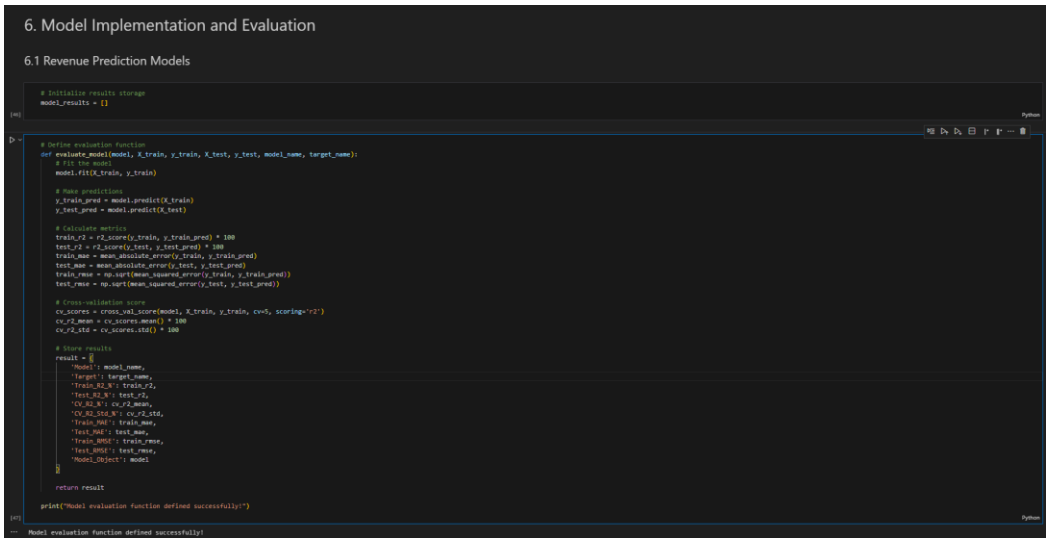

Figure 12: Model Building

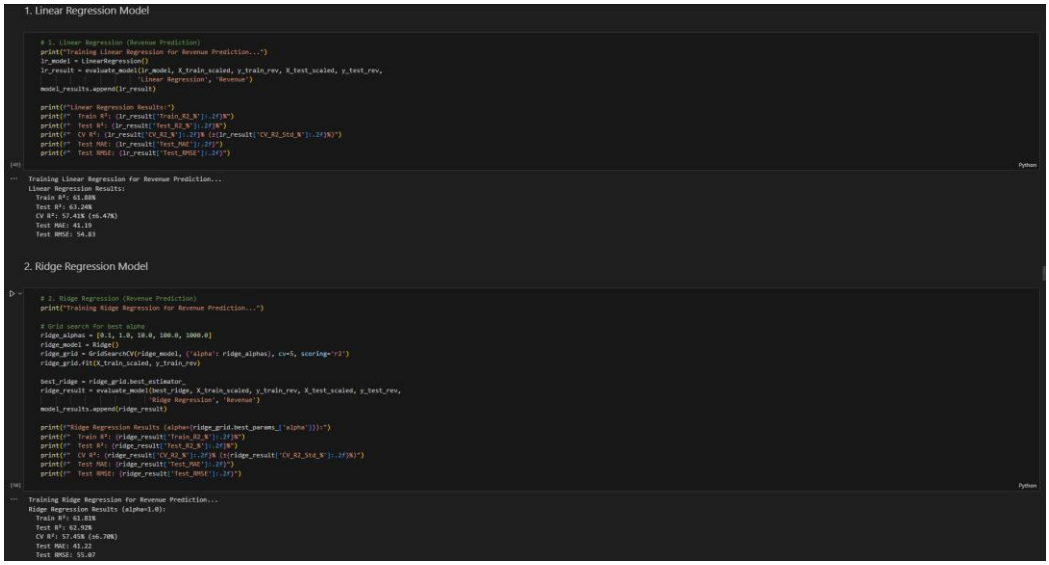

Figure 13: Linear and Ridge Regression Models

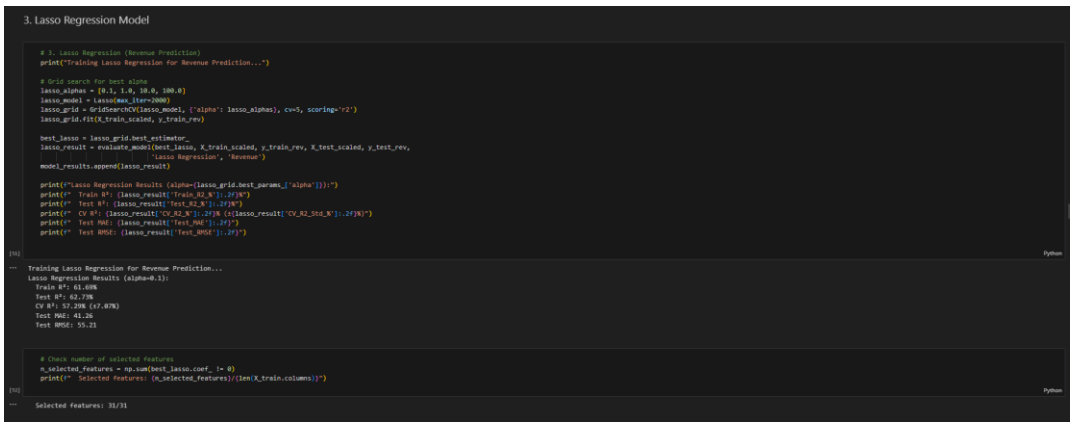

Figure 14: Lasso Regression Model

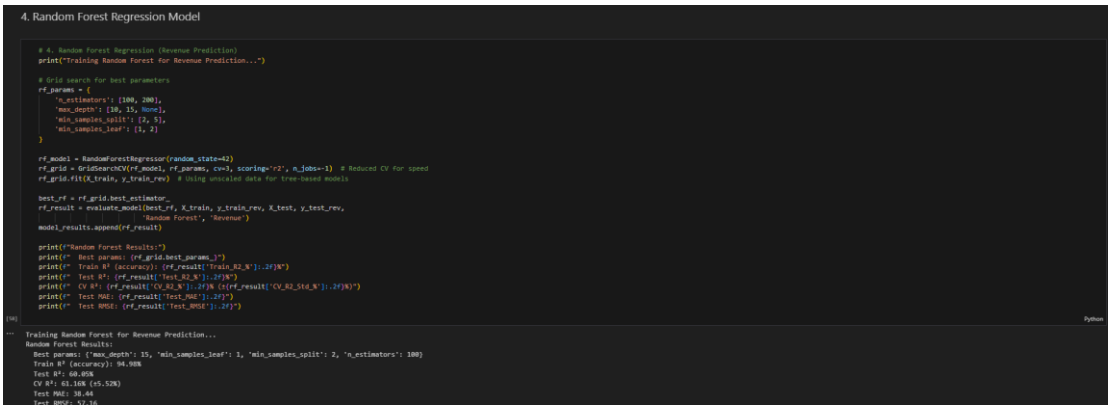

Figure 15: Random Forest Model (Tree Based Model)

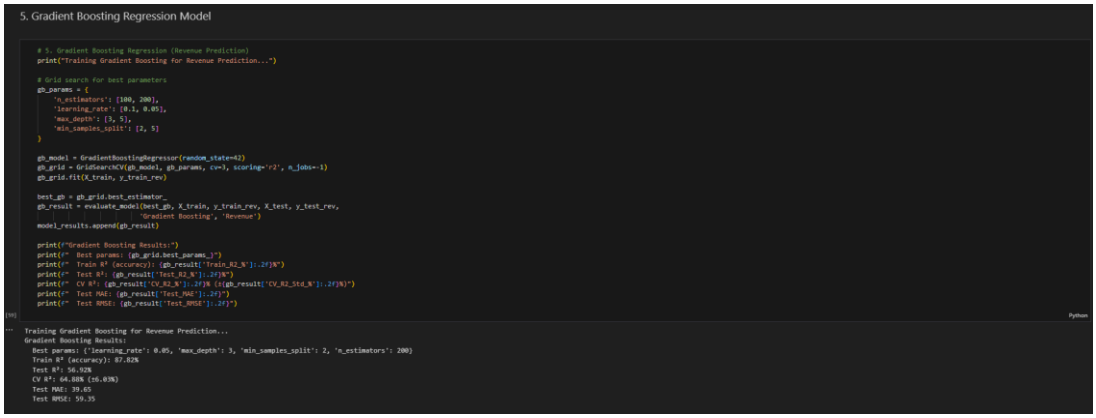

Figure 16: Gradient Boosting Model

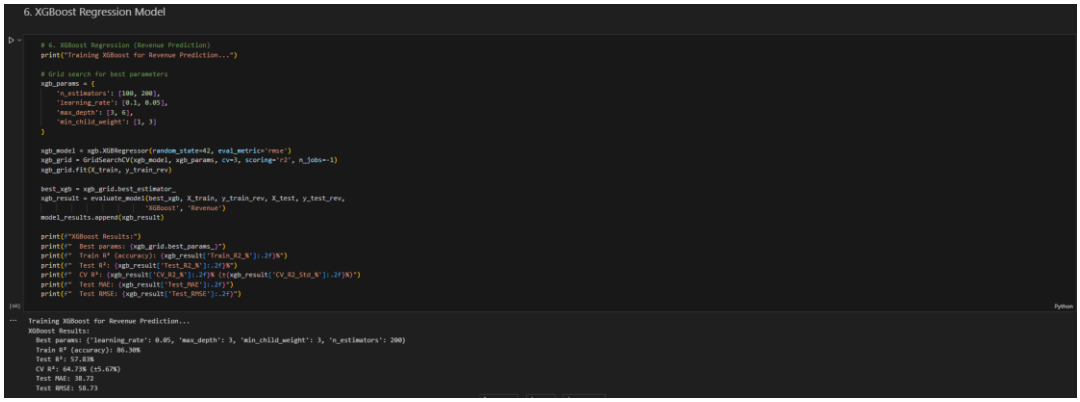

Figure 17: XGBoost Model

2. **Rating Prediction**:

- Feature matrix: 30 columns (exclude Rating).
- Target: y_rating = df_processed['Rating'].

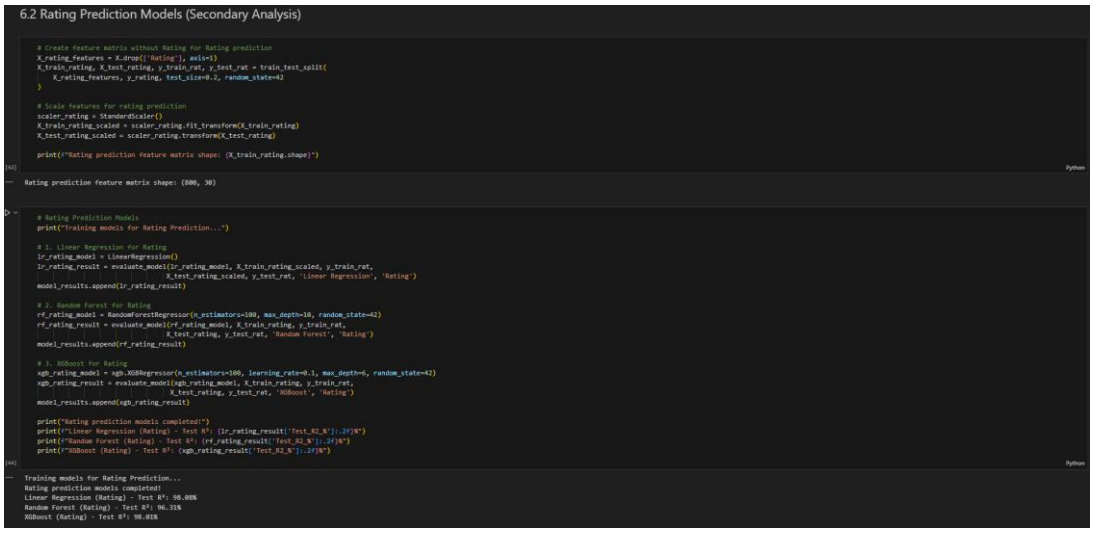

Figure 18: Secondary analysis Rating prediction models

6 Model Comparisons

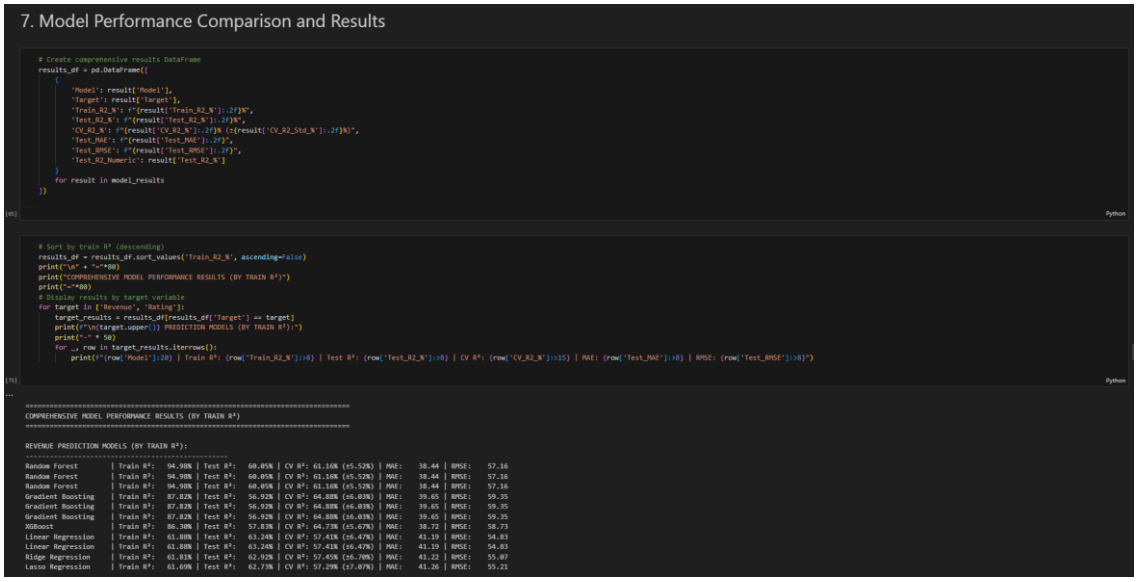

Figure 19: Model Comparison

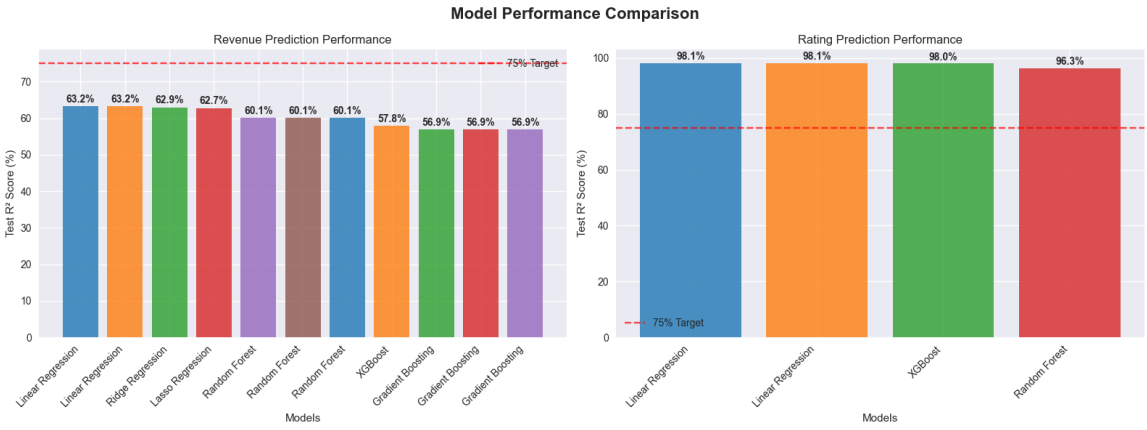

Figure 20: Plots of Model comparison

Figure 21: Feature Importance

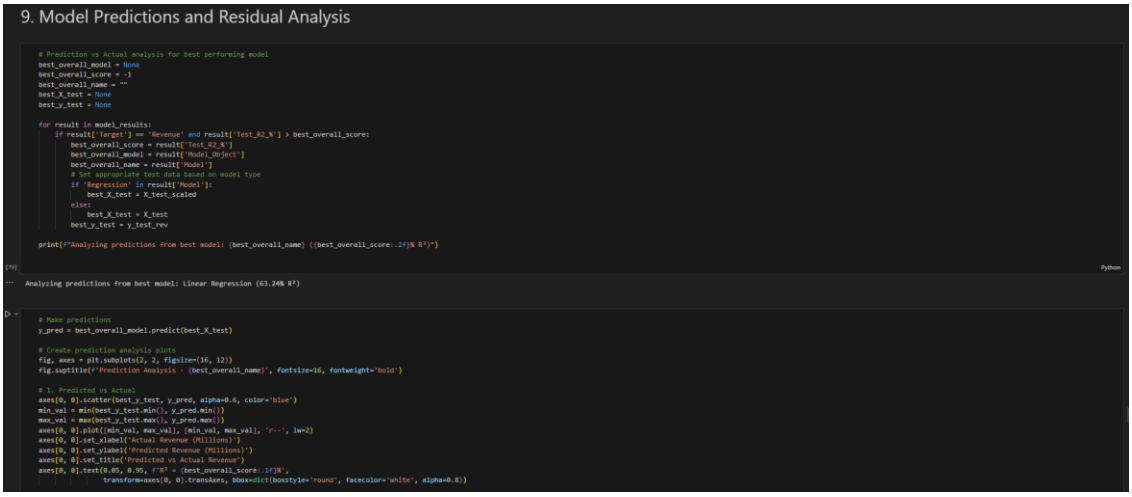

Figure 22: Model Prediction and residual analysis

Figure 23: Business Insight and Recommendation

7 Troubleshooting

- **Missing Dependencies**: Ensure all libraries are installed with correct versions (pip list).
- **Dataset Issues**: Verify imdb_movie_dataset.csv is in thesis_data/ and matches the expected structure.
- **Memory Errors**: Reduce n_estimators (e.g., to 500) for Random Forest/XGBoost if RAM is limited.
- **Convergence Warnings**: Increase max_iter for Ridge/Lasso (e.g., Ridge(max_iter=10000)).
- **File Output Errors**: Check write permissions in thesis_data/ directory.

References

GeeksforGeeks (2021). *Matplotlib Tutorial*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/python/matplotlib-tutorial/.

Lee, S. (2025). *Mastering Advanced Regression Techniques for Improved Data Insights*. [online] Numberanalytics.com. Available at: https://www.numberanalytics.com/blog/mastering-advanced-regression-techniques-for-improved-data-insights.

Matplotlib (2024). *Matplotlib: Python plotting — Matplotlib 3.3.4 documentation*. [online] matplotlib.org. Available at: https://matplotlib.org/stable/index.html.

Mohini Gore, Aishwarya Sheth, Samrudhi Abbad, Paryul Jain and Prof. Pooja Mishra (2022). *IMDB Box Office Prediction Using Machine Learning Algorithms*. [online] Ijraset.com. Available at: https://www.ijraset.com/research-paper/imdb-box-office-prediction-using-ml-algorithms.

Pandas (2024). *pandas documentation — pandas 1.0.1 documentation*. [online] pandas.pydata.org. Available at: https://pandas.pydata.org/docs/.

ResearchGate. (n.d.). *(PDF) Predicting Movie Success Based on IMDB Data*. [online] Available at: https://www.researchgate.net/publication/282133920_Predicting_Movie_Success_Based_on_IMDB_Data.

Sanyal, A. (2025). *Predicting Movie Success Based on Imdb Data*. [online] Scribd. Available at: https://www.scribd.com/document/431502642/Predicting-Movie-Success-Based-on-Imdb-Data.

Scikit-learn (2019). *scikit-learn: machine learning in Python — scikit-learn 0.20.3 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/index.html.

seaborn (2012). *seaborn: statistical data visualization — seaborn 0.9.0 documentation*. [online] Pydata.org. Available at: https://seaborn.pydata.org/.

Zhang, Z., Meng, Y. and Xiao, D. (2024). Prediction techniques of movie box office using neural networks and emotional mining. *Scientific Reports*, 14(1). doi: https://doi.org/10.1038/s41598-024-72340-z.