Detailed Project Report
Ensemble Model Development and Visualization for Predicting Average Number of Receipts for 2022

GitHub Link
Predict Average Receipts Here

Introduction

This project involves development of an ensemble machine learning model to predict number of Receipts received per day in 2022 and the subsequent visualization of its predictions. The goal is to take an input from user which is a month of 2022 and the predict the total number of Receipts for that month. The ensemble model was created by integrating predictions from three different models: XGBoost, Neural Network, and Linear Regression. The final output was visualized through a bar graph embedded to a custom-styled HTML file. The html file is then stored in an AWS S3 bucket and make it public so that it is accessible to everyone.

Objective

The primary objective was to harness the predictive power of three distinct models, thereby improving the overall prediction accuracy through ensemble techniques. The secondary objective was to visually present these predictions in an easily interpretable format, using web technologies for accessibility and user interaction.

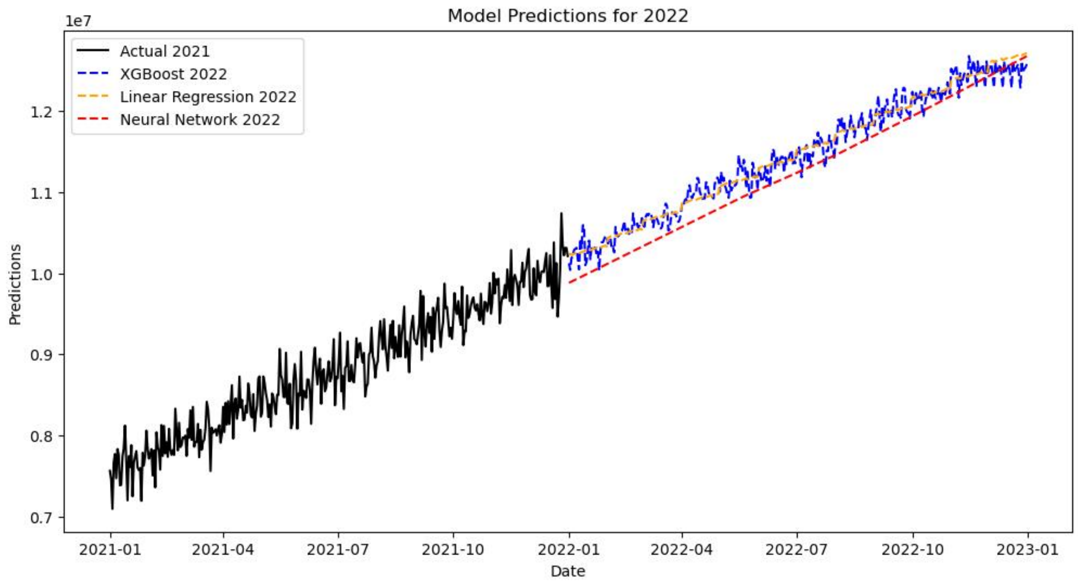
Process Overview

1. Data Acquisition and Preprocessing:
   - Source: The data, comprising dates and receipt counts, was loaded from an Excel file.
   - Preprocessing: The data underwent preprocessing, including conversion of dates and extraction of additional features like day of the week, month, and day of the year to understand the patterns involved in the data.

2. Model Development:
   - XGBoost Model: Utilized for its performance in regression tasks. Individual XGBoost outperforms other models. Trained on features including the day of the week, month, and day of the year.
   - Neural Network Model: Implemented in PyTorch.
     - Focused on 'DayOfYear' as the primary feature. Receipt counts were normalized to aid in the model's performance.

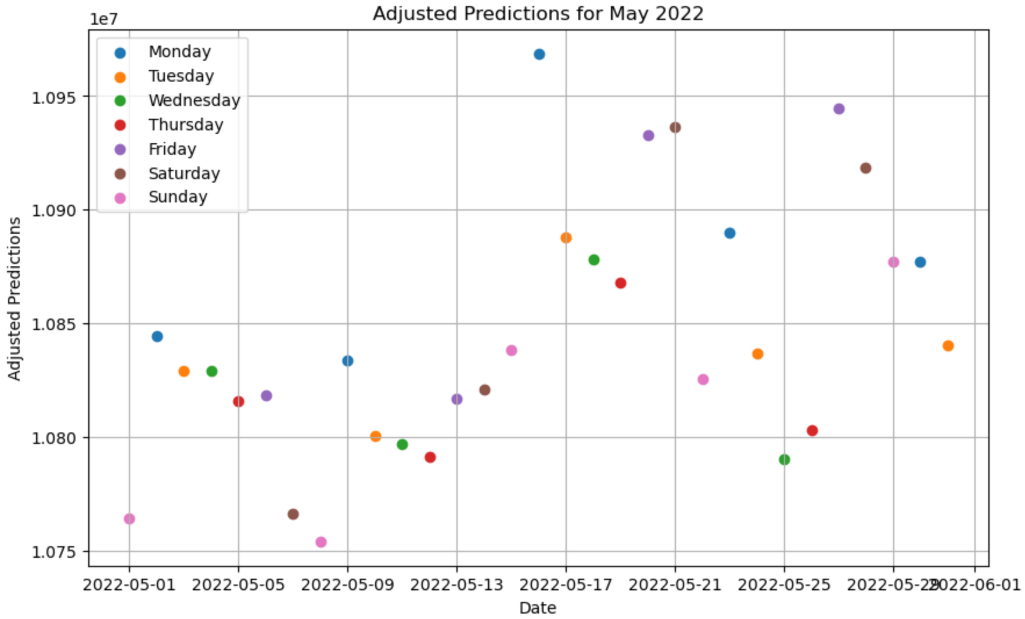
Linear Regression Model: Employed a custom implementation using gradient descent. Allowed for a fundamental approach to predicting trends.

3. Ensemble Technique:
   - Combined the individual predictions from each model by calculating their average.
   - Ensured that the model capitalized on the strengths of each individual model while mitigating their weaknesses.

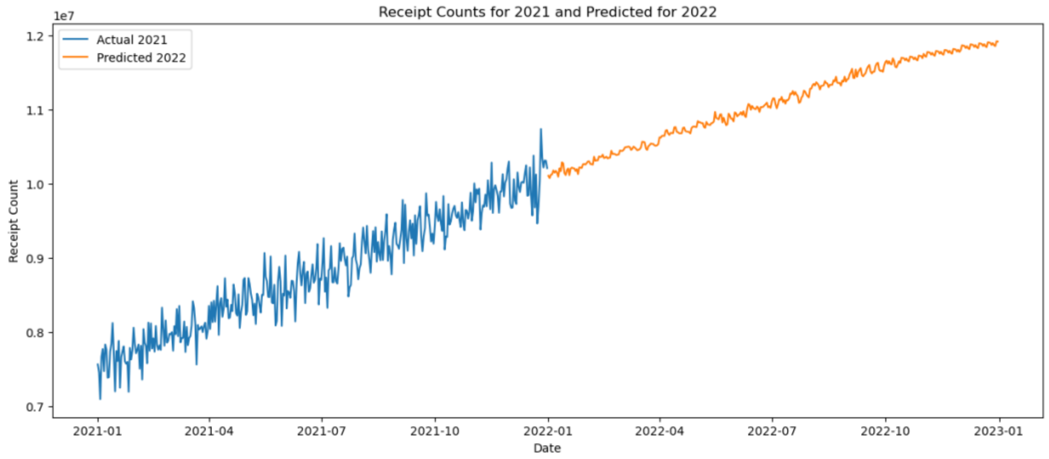


4. Prediction Adjustment:
- Adjusted the ensemble predictions for 2022 by adding the difference between the values on December 31, 2021, and January 1, 2021, addressing potential continuity issues in the data.

5. Visualization:
- Generated a line graph using Matplotlib to depict the ensemble predictions. Created a scatterplot. Each day of the week is assigned a different color to identify patterns in the receipts received.



- Saved this graph as an image to be embedded in the HTML file.

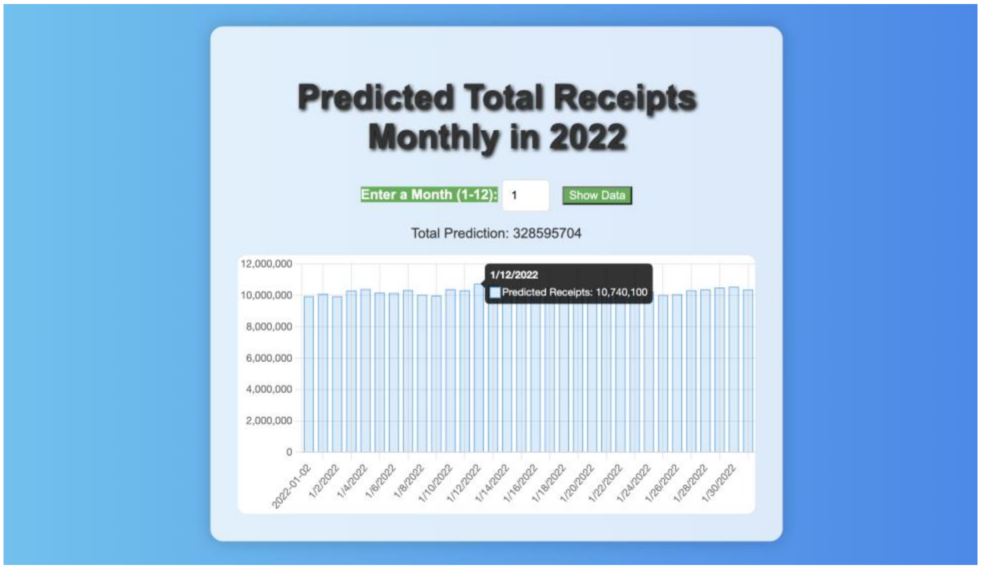


HTML File Development

As the model has already predicted all the receipts for 2022, these predictions are stored in an xlsx file and then used for creating an HTML file.

1. Basic Structure and Styling:
   - Created an initial HTML layout with input elements for user interaction.
   - Applied CSS to enhance the visual appeal, including a gradient background and styled text.

2. Graph Integration and Final Styling:
   - Embedded the line graph image within the HTML file.
   - Adjusted the size and styling of the graph to align with the overall design of the page.
   - Made final adjustments to the layout and elements for a cleaner and more elegant appearance.
   - When user hovers to a particular day of month the predicted receipts can be seen.



Predicted Total Receipts Monthly in 2022

Enter a Month (1-12): 1

Total Prediction: 328595704

Finally, the HTML file is stored in Amazon S3 bucket and made it publicly available for everyone.

Conclusion

The project successfully demonstrated the creation of an ensemble model that combines the predictive capabilities of three different machine learning algorithms. The visualization provided a clear and aesthetically pleasing presentation of the model's year-long predictions. This project exemplifies how machine learning can be effectively coupled with web technologies for data presentation, with certain limitations in interactivity due to the constraints on scripting.