CS 59000-011, Fall 2024

Purdue University Northwest

11/14/2024

Name	Email
Shivam Pandya	pandya23@pnw.edu

Table of Content

ABSTRACT

This project, developed in collaboration with No Limit Living, focuses on designing a predictive model that analyzes key metrics for assessing property viability and forecasting area development trends. Our objective was to deliver a comprehensive, data-driven tool to help No Limit Living make informed decisions about property investments, guided by future market projections. We began with a foundational understanding of the MERN stack, preparing for potential implementation, but eventually transitioned to using Flask for streamlined model deployment. Our team collected and curated high-quality datasets from official government sources throughout the project, ensuring the data was comprehensive and aligned with project goals. Various machine learning models, including Random Forest, XGBoost, and Linear Regression, were tested and evaluated; however, after rigorous experimentation, we determined that time series modeling with the ARIMA and auto-ARIMA frameworks offered the most accurate predictions for the long-term statistical analysis that No Limit Living required. This documentation provides a detailed overview of our development journey, highlighting our model selection process, data acquisition efforts, and the predictive insights generated for No Limit Living's strategic property decisions.

I. INTRODUCTION

The project undertaken with No Limit Living was aimed at developing a predictive analytics tool to assist the company in making strategic property investment decisions. No Limit Living specializes in real estate solutions, emphasizing data-informed insights to maximize investment potential and manage risk effectively. Our task was to create a model that could predict future trends in property values, population growth, economic factors, and other vital metrics essential for evaluating the attractiveness and sustainability of a given area.

Our project journey began with learning the MERN stack to understand full-stack web development fundamentals, as it was initially considered for implementing the final model. However, we ultimately decided on Flask for the deployment due to its flexibility and compatibility with the data models we created. Data gathering was a crucial step, as we sought datasets from verified government sources to ensure reliability. Where data fell short of perfection, we refined and adjusted it to align with project requirements.

The model development process involved testing multiple algorithms, including Random Forest, XGBoost, and Linear Regression. Through iterative testing, we identified that time series models—specifically ARIMA and auto-ARIMA—were best suited to capture trends over time and provide accurate forecasts. This report details our methodology, from data acquisition to model selection, ultimately showcasing how our tool equips No Limit Living with actionable insights to make informed, data-backed property investment decisions.

II. DATA COLLECTION AND FINE-TUNING

After extensive research and web scraping, we identified several key data sources that offered comprehensive economic, demographic, and real estate information necessary for our model. Among these sources, two websites provided the most valuable datasets: the Bureau of Economic Analysis (BEA) Interactive Data Tables (bea.gov) and the Federal Reserve Economic Data (FRED) platform (fred.stlouisfed.org). These platforms offered detailed historical data across various economic indicators, which are critical for accurately forecasting future trends for real estate investments.

The datasets we accessed were comprehensive but unrefined, requiring significant filtering and transformation to make them usable. We carefully selected relevant metrics, such as income levels, employment rates, population statistics, and housing market indicators, to create a dataset that aligned with the project's predictive goals. This structured dataset, shown in the example below, served as the foundation for model training and analysis.

The process of transforming this raw data into an "ultimate dataset" involved several rounds of data cleaning and formatting adjustments to ensure compatibility with our predictive models. The cleaned and consolidated dataset not only provided a solid basis for our initial analyses but also allowed for greater accuracy in forecasting future property trends for No Limit Living.

GeoName	Year	All industry total	Personal income (thousands of dollars)	Population (persons) 3/	Total employment	Wages and salaries
Adams, IN	2010	0.011769621	0.020184287	0.031550827	0.02418583	0.013207965
Allen, IN	2010	0.188604976	0.322695572	0.389373501	0.320691209	0.25371683
Bartholomew, IN	2010	0.062878551	0.071691395	0.078648423	0.071662264	0.066275948
Benton, IN	2010	0.002263287	0.002048843	0.003088659	0.002693041	0.001415337
Blackford, IN	2010	0.001750753	0.00462814	0.007446827	0.004041769	0.002463492
Boone, IN	2010	0.022118246	0.07259287	0.056605012	0.041697219	0.026643108
Brown, IN	2010	0.001246372	0.008230661	0.010161643	0.004941413	0.001196648
Carroll, IN	2010	0.005426933	0.011928771	0.015705874	0.008098263	0.004284073
Cass, IN	2010	0.012828539	0.025288566	0.036616628	0.02473504	0.015977775
Clark, IN	2010	0.050329436	0.095205007	0.116306471	0.085860728	0.060772355
Clay, IN	2010	0.00593192	0.015758743	0.023122661	0.011573154	0.005937919

Figure 1: Refined Dataset

III. MODEL CREATION

The journey of model creation began with uncertainty, as I explored which predictive model would best suit our dataset and provide meaningful results. Starting with familiar machine learning models like **Linear Regression**, **Random Forest**, and **XGBoost**, I quickly realized that while these models are powerful, they failed to yield informative or accurate predictions for our dataset. Initially, I suspected the dataset might be the issue, so I revisited and refined it, but the results remained suboptimal.

While researching potential reasons for the lack of performance, I discovered a category of models specifically designed for **time series analysis**. These models are well-suited for datasets where trends and patterns evolve, offering the capability to forecast future values based on historical data. Excited by this insight, I started exploring various time-series models.

My first attempt was with **Long Short-Term Memory (LSTM) networks**, a type of recurrent neural network designed for sequential data. However, the results did not meet expectations,

likely due to the complexity of tuning such models for our specific use case. Continuing my research, I came across **ARIMA (AutoRegressive Integrated Moving Average)** and its extension, **auto_arima**, which automates selecting the best parameters for a given dataset.

Implementing **auto_arima** proved to be a turning point. This model not only predicted future values with high accuracy but also passed the **Jarque-Bera test**, indicating how well the model fits the data and whether the residuals followed a normal distribution. After a long process of trial and error, supported by extensive research and testing, I arrived at a time-series model that consistently delivered reliable and actionable results for forecasting. This experience highlighted the importance of persistence and adaptability in model selection and optimization.

```
Jarque-Bera test results for Allen, IN:
  - JB Stat: 0.3345596265520736
  - p-value: 0.8459628665754673

Jarque-Bera test results for Bartholomew, IN:
  - JB Stat: 0.49321008315679327
  - p-value: 0.7814492725606756

Jarque-Bera test results for Benton, IN:
  - JB Stat: 2.6144361629092527
  - p-value: 0.27057171885057

Jarque-Bera test results for Blackford, IN:
  - JB Stat: 0.4783240448021117
  - p-value: 0.7872873138575764
```

Figure 2: Jarque-Bera Test

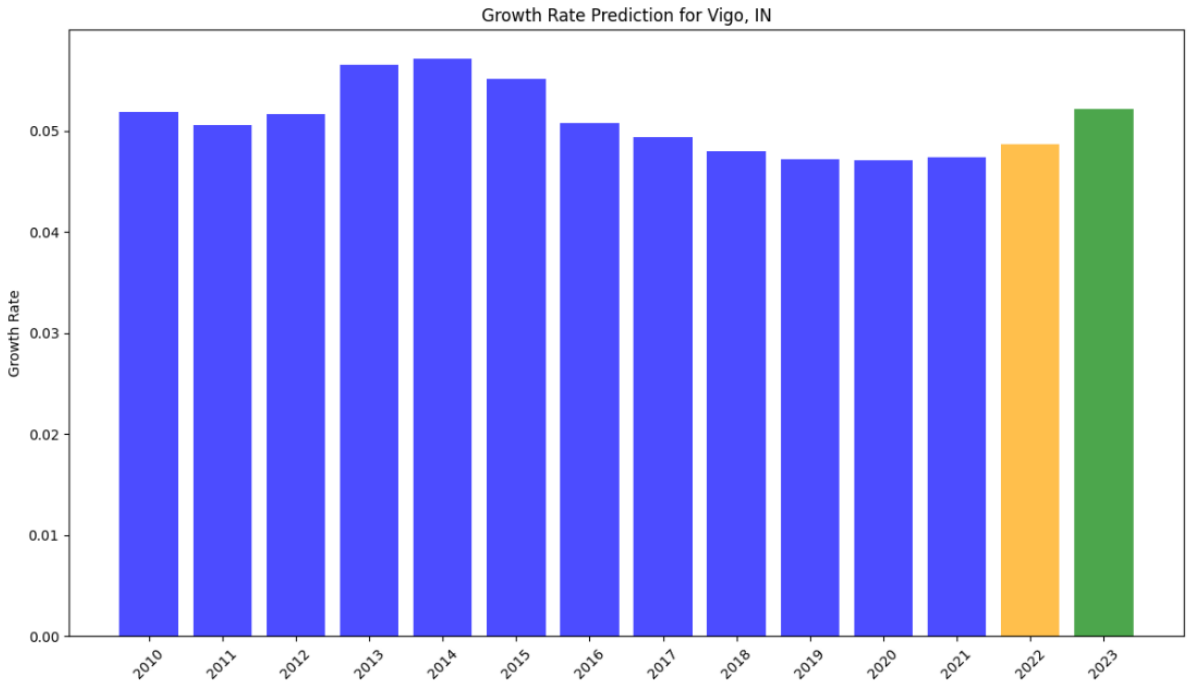

Figure 3: Predicting 2022 and 2023 percentage growth

IV. USER INTERFACE

Initially, our goal was to develop the User Interface (UI) using the **MERN stack** as planned. However, due to time constraints this semester, we decided to postpone the full MERN-based UI implementation to the next semester and opted to create a basic UI using **Flask** to complement our backend.

For the initial design, we focused on simplicity and functionality. We created **PKL files** containing the pre-processed models for all the counties and integrated them into the backend. The UI allowed users to input the name of a county into a text field, and upon pressing a button, it displayed the predicted percentage growth for future years. This design served as a functional prototype, effectively showcasing the backend's predictive capabilities.

As the project progressed, we enhanced the UI by introducing an interactive **map feature**. In this improved design, users could click on a specific county on the map to view detailed metrics on the side panel, including predictions for future years. This upgrade not only improved the usability and visual appeal of the interface but also provided an intuitive way for users to interact with the data. These iterative improvements highlight our commitment to creating a seamless and user-friendly experience, laying the groundwork for a more robust UI in the next semester.

Figure 4: Initial UI with a simple layout

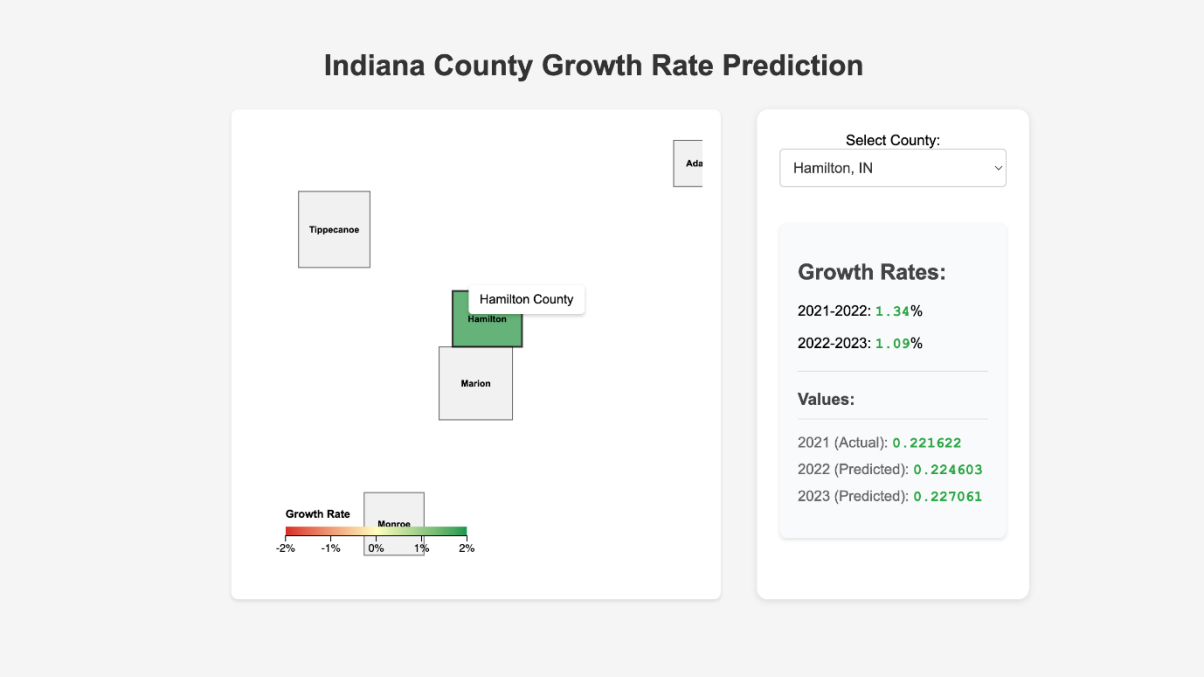

Figure 5: Map feature in the UI

V. DISCUSSION

The project undertaken this semester reflects the culmination of our efforts in data collection, model development, and interface design to address the practical challenges posed by No Limit Living. The primary objective was to create a predictive model that provides actionable insights into the growth metrics of various counties, assisting in property investment decisions. This project journey has been both educational and transformative, highlighting the complexities of real-world data analysis and predictive modeling.

One of the most significant challenges we faced was identifying the right model for the dataset. While starting with traditional machine learning models such as Linear Regression, Random Forest, and XGBoost, their outputs proved inadequate in predicting future trends effectively. This led us to explore time-series models, where we eventually implemented ARIMA with Auto-ARIMA for parameter optimization. This approach demonstrated the importance of aligning the model type with the nature of the data, showcasing a clear learning curve in understanding advanced predictive modeling techniques.

Another key aspect of the project was data acquisition and preparation. Gathering data from official sources such as the **Bureau of Economic Analysis (BEA)** and **FRED** required extensive effort in filtering and fine-tuning the datasets to make them usable. The process underscored the importance of accurate data preprocessing in ensuring the reliability of model outcomes. The iterative process of refining datasets and models highlighted the interdependence between data quality and model performance.

In terms of user engagement, the development of the interface provided insights into translating technical outputs into an accessible and interactive format. The transition from a simple Flask-based text input UI to an interactive map-based feature demonstrates the potential of integrating usability with advanced analytics. Although time constraints limited the scope of the UI development, the foundational work lays the groundwork for further enhancements, which will be pursued in the next semester.

This project has provided invaluable lessons in balancing theoretical understanding with practical application. By combining our technical expertise with real-world problem-solving, we not only achieved significant milestones but also identified areas for further improvement. Looking ahead, the implementation of a MERN-based UI and scaling the model for broader use cases are the next steps in continuing this impactful collaboration.

VI. CHALLENGES

This project presented several challenges, starting with data acquisition and cleaning. While sourcing data from official websites like the **Bureau of Economic Analysis (BEA)** and **FRED**, we found that the datasets were often incomplete or unfiltered, requiring extensive cleaning and fine-tuning. Another significant challenge was selecting the right predictive model. Initially, popular models like Linear Regression, Random Forest, and XGBoost failed to yield meaningful results, leading us to explore time-series analysis and eventually adopt the ARIMA model. Time constraints during the semester also impacted our progress, as we had to postpone the planned MERN stack-based UI to the next semester and instead focus on a functional Flask-based interface. Additionally, computational limitations hindered our ability to experiment with resource-intensive models like Long Short-Term Memory (LSTM) networks. Finally,

integrating advanced features such as map-based interactivity posed technical hurdles in ensuring seamless communication between the back-end model and front-end visualization. Despite these obstacles, each challenge provided valuable lessons and insights, fostering growth and innovation throughout the project.

VII. CONCLUSION

This project was a comprehensive learning experience that combined skills in data collection, model creation, and user interface development to address a real-world problem for No Limit Living. By exploring diverse datasets and experimenting with various predictive models, we ultimately developed a time-series-based solution using the ARIMA model that provides reliable growth forecasts for counties across the United States. Although we faced challenges, such as data preprocessing, model selection, and UI implementation, these obstacles enhanced our problem-solving abilities and teamwork. The Flask-based interface served as a stepping stone, offering a functional solution while laying the groundwork for a more advanced MERN-based UI in the future. Overall, this project not only met its objectives but also prepared us to tackle more complex challenges in data-driven decision-making and predictive modeling.

VIII. ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to **No Limit Living** for providing us with the opportunity to work on this project and contribute to their mission of aiding informed property decisions. We are deeply grateful to our course instructors and mentors for their invaluable guidance and feedback throughout the semester, helping us navigate challenges and improve our skills. Additionally, we thank official data sources such as **BEA.gov** and **FRED** for providing us with reliable datasets that formed the foundation of our analysis. Finally, we extend our appreciation to our teammates for their collaboration, dedication, and relentless efforts, which made this project a success. This experience has been both enlightening and rewarding, enabling us to grow as professionals and problem-solvers.