Forecasting US County Unemployment

Team Members

- Saloni Nitin Bonde

- Ramyalakshmi Lakshmi Narasimhan

- Subbareddy Bhumireddy Venkata

- Krishnam Malhotra

Faculty: Prof. Adem Orsdemir

Course: MGT 256 - Business Analytics for Management

Acknowledgments

We express our sincere gratitude to Dr. Orsdemir for his invaluable guidance and unwavering support throughout the project. His expertise and mentorship played a crucial role in the successful completion of this endeavor.

Table of Contents **Page No**

Executive Summary

This report focuses on predicting annual unemployment across all U.S. counties for 2022, utilizing data from 2021. Our objective is to enhance business analytics methodologies for informed decision-making in diverse economic landscapes. We explore the problem statement, goal, and why it is important for policymakers, economic planners, and businesses.

Introduction

This project focuses on predicting annual unemployment by leveraging historical data, emphasizing its pivotal role in informed decision-making across sectors. In the dynamic landscape of today's economies, understanding labor market trends holds significance for policymakers, businesses, educational institutions, and economists alike. For policymakers, historical data aids in crafting effective employment policies and strategies. Businesses benefit by using labor market insights to make informed decisions on hiring and expansion. Educational institutions tailor programs to meet local job market needs. Economists rely on this data for accurate economic forecasting. Through comprehensive data analysis, this project aims to empower stakeholders with insights crucial for policy development, economic planning, workforce development, and economic forecasting, contributing to the resilience and adaptability of economies.

Overview

In this project, we harnessed the power of libraries such as tidycensus, tidyverse, xlsx, caret, and applied analytical methods like KNN analysis and Linear regression. Our meticulous process began with comprehensive Data Collection, involving diverse sources, types, and rigorous

sampling. Data Cleaning and Preprocessing addressed missing data and outliers, ensuring robust datasets. Through Exploratory Data Analysis (EDA), we delved into descriptive statistics, data visualization, and correlation analysis. Model Training included data splitting and parameter tuning for optimal results. Finally, Model Evaluation employed performance metrics, offering a comprehensive and insightful analysis of our findings in this multifaceted journey of data exploration and analysis.

Choosing a Dataset

Given the broad scope of our problem statement, selecting relevant datasets was paramount. This proved challenging as pertinent data was dispersed across various sources. We diligently curated information from diverse platforms, including the United States Department of Agriculture (https://www.ers.usda.gov/data-products/county-level-data-sets/), the United States Census Bureau (https://data.census.gov/),United States Department of Health & Human Services (https://catalog.data.gov/organization/hhs-gov), United States Bureau of Economic Analysis (BEA) (https://www.bea.gov/), and the Economic Policy Institute (https://www.epi.org/resources/budget/).

Accessing census data required familiarity with tidycensus, facilitating the extraction of demographic and industry employment patterns. The complexity arose in unifying this disparate data, a challenge addressed through Streamlining Data Integration: Unifying Files with FIPS Codes. This comprehensive approach ensured a holistic consideration of essential datasets, providing a solid foundation for subsequent analyses and insights in our pursuit of addressing the intricacies of the broad problem statement.

Research Problem

Utilize business analytics methodologies by predicting annual unemployment across U.S. counties, utilizing historical data. This research addresses the imperative need for informed decision-making in diverse economic landscapes, contributing to a comprehensive understanding of labor market dynamics and fostering strategic planning for businesses nationwide.

Predictors

We delve into the pivotal role of predictors. The meticulous process of identifying and selecting predictors is essential to address our research problems effectively and to attain precise predictions. The following variables were scrutinized for their potential impact: "People 0-17 in Poverty '21," "Labor Force in Poverty," "Bachelor's or Higher," "GDP '21," "Civilian Labor Force '21," "Pop. Estimate," "Pop. Change '21," "Births '21," "Deaths '21," "Net Int'l Migration," "Net Domestic Migration," "Residual," "Female/Male Estimate," "Agri./Finance/Info./Manuf. Estimate," "Median Family Income," "Taxes," "Rent," "Food," "Health Care," "Violent Crime Rate," "Smokers," and "Teen Birth Rate."

Handling Missing Values

In this section, we delve into the crucial aspect of managing missing values within our dataset. We examine diverse approaches to address the absence of data, thereby fortifying the reliability of our analysis. Our initial decision involved excluding counties for which we lacked data. Subsequently, we moved on to the Data Cleaning phase, where we judiciously addressed missing values or observations through appropriate imputation or removal. This meticulous process aims to ensure the coherence of our dataset.

Key Steps:

1. Standardizing Formats: Guarantee consistency in data formats and units to enhance the overall integrity of the dataset.

2. Handling Outliers: Evaluate and address extreme values as necessary, fostering a more accurate and reliable analysis.

3. Correlation Validation: Conducted a correlation analysis to gain valuable insights and thoroughly explored variable relationships through correlation assessments.

Dataset Size

Post data cleaning, our dataset comprised of 3,143 counties, encompassing 36 variables and 96,588 datapoints. To streamline our analysis, we selected a representative sample of 1,000 counties using random sampling techniques, ensuring a focused and statistically sound approach to our research.

Reducing Number of Categories

Streamlining interlinked datasets, we opted to reduce variables to 25 for enhanced focus and clarity. This decision ensures a more concise and manageable dataset, facilitating a streamlined analysis while preserving the essential interconnected nature of the retained variables.

Data Exploration/Visualization

After looking at the mean, median, and standard deviation we draw the following conclusions.

1. Educational Attainment: "Bachelor_degree_or_higher_2017.21" exhibit varying degrees of educational achievement, with mean values of 21,552.65 and 25,644.65, respectively.

2. GDP and Economic Output: The "GDP_2021" variable indicates a wide range of economic outputs, with a mean of 8,924,031. This suggests substantial economic disparities among the sampled regions, which could be influenced by factors such as industry composition and regional development.

3. Poverty Estimates: The "Estimate_of_people_of_all_ages_in_poverty_2021" variable suggests that there is considerable variation in poverty levels across the regions, with a mean of 13,474.56. This underscores the importance of understanding and addressing economic disparities and potential social challenges.

4. Migration Patterns: The migration-related variables ("Net_international_migration," "Net_domestic_migration," "Net_migration") reveal interesting patterns. Positive values indicate net migration into the region, while negative values suggest net migration out. These figures can provide insights into regional attractiveness and economic opportunities.

5. Births and Deaths: The "Number_of_people_born_in_2021" and "Number_of_people_died_in_2021" variables highlight the demographic dynamics within the regions. The positive net births contribute to population growth, while deaths represent a factor influencing population stability.

6. Economic Sector Contributions: Variables such as "Estimate_Agriculture," "estimate_Finance," "Estimate_of_Information," and "Estimate_of_Manufacturing" provide insights into the economic structure of the regions. Understanding the dominant sectors can inform economic development strategies.

7. Median Family Income and Living Costs:

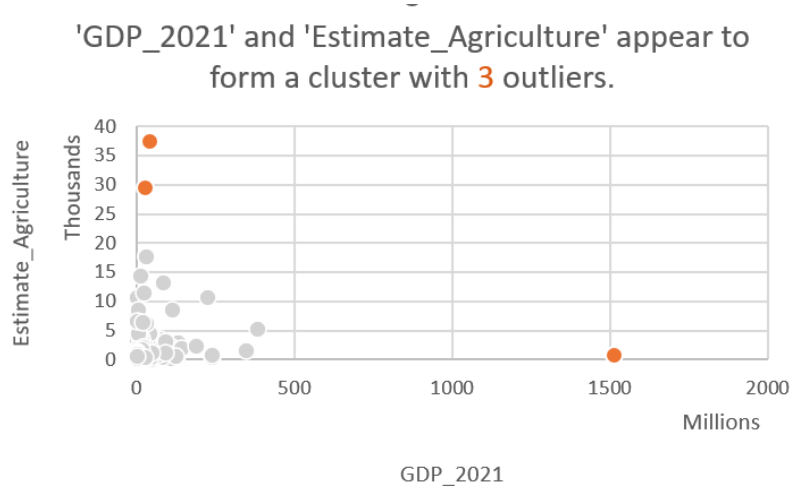
The "Median_family_income" variable indicates a median income of 65,400.60. Comparing this with the cost-related variables such as "Taxes," "Rent," "Food," and "Health_care" can shed light on the affordability and standard of living in the regions.

8. Social and Health Indicators: Variables like "Violent_Crime_Rate," "Smokers," and "Teen_Birth_Rate" provide insights into social and health challenges. Understanding these indicators is crucial for designing effective public health and safety interventions.

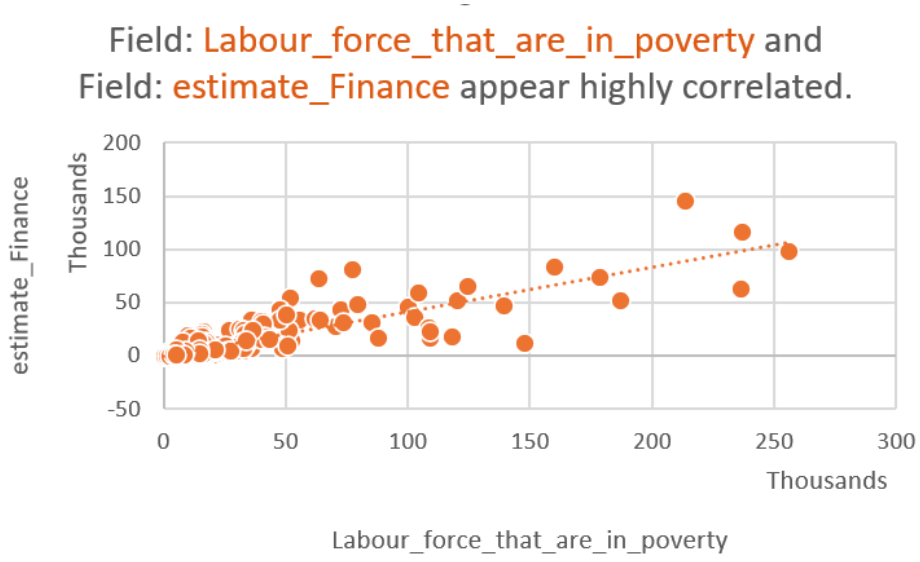
9. Temporal Trends: The variables "month" and "day" can be used to analyze temporal patterns. Examining trends over time may reveal seasonality or specific temporal factors affecting the regions.

In conclusion, this analysis provides a multifaceted view of the sampled regions, encompassing education, economic indicators, demographics, migration, employment, and health. The insights derived from these variables can guide policymakers, researchers, and community leaders in developing targeted strategies to address challenges and promote sustainable development across the regions.

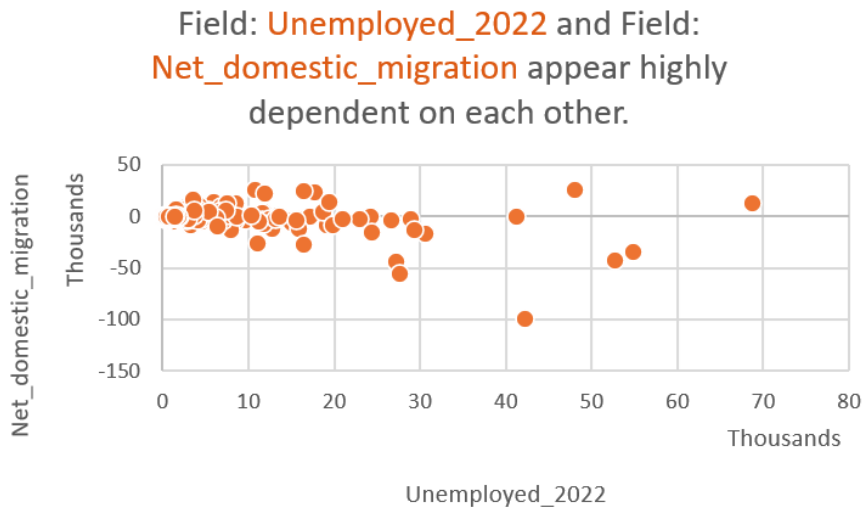
Data Visualization:



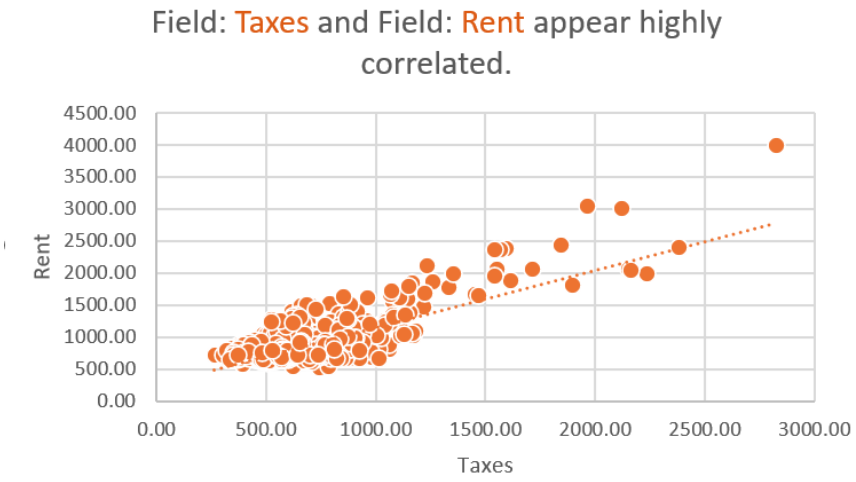
The variables 'GDP_2021' and 'Estimate_Agriculture' exhibit a noticeable clustering pattern in the data, suggesting a potential relationship. Three outliers stand out within this cluster, indicating regions that deviate significantly from the general trend. This clustering may imply a shared economic characteristic or regional specialization in agriculture. The outliers, while representing exceptions, could indicate areas with unique economic structures or challenges.



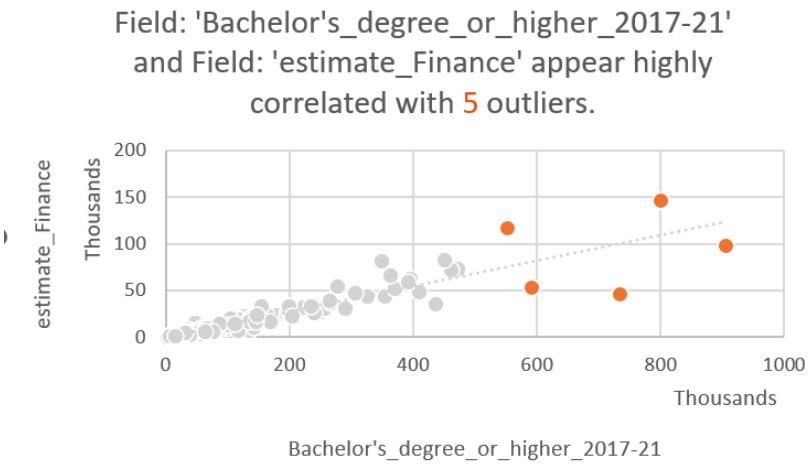
The variables 'Labour_force_that_are_in_poverty' and 'estimate_Finance' exhibit a strong positive correlation, indicating a significant relationship between the labor force in poverty and financial sector estimates. This correlation suggests that regions with higher poverty rates tend to have lower financial sector estimates or vice versa. The link may stem from economic disparities, where areas facing financial challenges also experience higher levels of poverty in the labor force.



The variables 'Unemployed_2022' and 'Net_domestic_migration' demonstrate a strong dependency, implying that unemployment rates are closely linked to domestic migration patterns. Higher unemployment rates may prompt individuals to migrate in search of employment opportunities, influencing the net domestic migration figures. Conversely, regions experiencing significant domestic migration may see fluctuations in their unemployment rates. This interdependence suggests a dynamic relationship between labor market conditions and population movements, emphasizing the impact of economic factors on migration decisions.

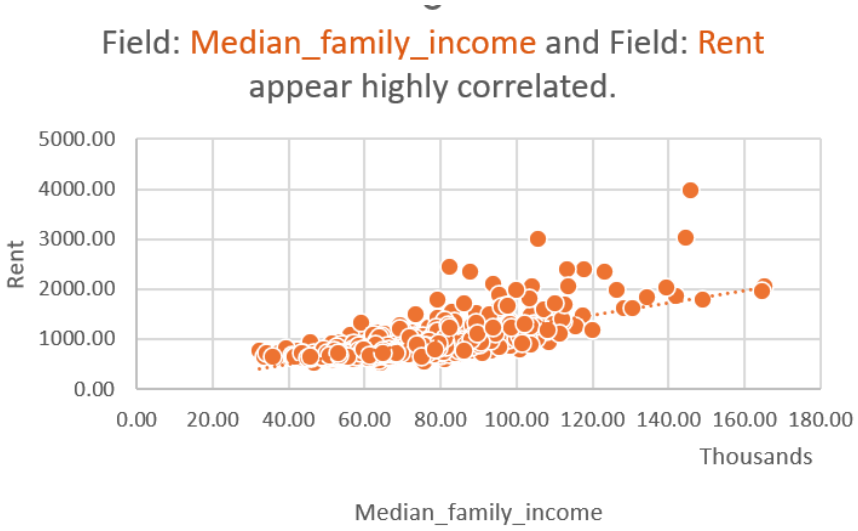


The strong correlation between 'Taxes' and 'Rent' indicates a notable relationship between tax burdens and rental costs. This suggests that regions with higher tax rates tend to experience elevated rental expenses or vice versa. The correlation underscores the intertwined nature of fiscal and housing policies, where tax structures may influence the affordability and pricing of rental properties.

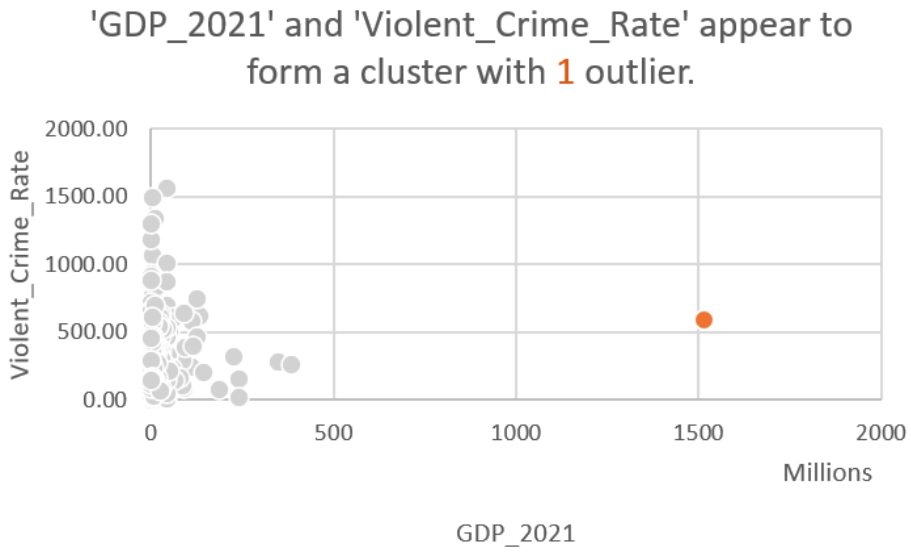


The strong correlation between 'Bachelor's_degree_or_higher_2017-21' and 'estimate_Finance' suggests a discernible relationship between the percentage of individuals with bachelor's degrees or higher and estimates in the finance sector. The presence of five outliers indicates regions

deviating significantly from the general correlation trend. This could signify unique economic landscapes where high educational attainment coincides with distinct patterns in the finance sector.



The robust correlation between 'Median_family_income' and 'Rent' suggests a pronounced relationship between median family income levels and rental costs. Regions with higher median family incomes tend to exhibit elevated rental expenses, indicating a correlation between economic prosperity and housing affordability. This connection underscores the impact of income disparities on housing affordability and emphasizes the challenges individuals in lower-income brackets may face in securing affordable rental accommodations.



The observed clustering of 'GDP_2021' and 'Violent_Crime_Rate' suggests a potential relationship between economic output and crime rates. However, the presence of one outlier indicates a region deviating significantly from this pattern. This outlier may represent a unique case where factors beyond economic indicators influence the crime rate.

Correlation-

	Unemployed_2022
High_school_diploma_only_2017.21	0.96
Some_college_or_associates_degree_2017.21	0.96
Bachelor_degree_or_higher_2017.21	0.90
GDP_2021	0.42
Labour_force_that_are_in_poverty	0.97
Unemployed_2022	1.00
Population_change_in_2021	-0.15
Number_of_people_born_in_2021	0.96
Number_of_people_died_in_2021	0.96
Net_international_migration	0.79
Net_domestic_migration	-0.35
Net_migration	-0.28
Residual	-0.06
Estimate_of_female	0.96
Estimate_of_male	0.96
Estimate_Agriculture	0.49
estimate_Finance	0.90
Estimate_of_Information	0.81
Estimate_of_Manufacturing	0.84
Median_family_income	0.33
Taxes	0.30
Rent	0.53
Food	0.25
Health_care	-0.20
Violent_Crime_Rate	0.20
Smokers	-0.37
Teen_Birth_Rate	-0.19

The correlation table provides insights into the interdependencies between unemployment and various socio-economic indicators. It suggests that unemployment is strongly related to education levels, poverty rates, and labor force size, while other factors show more varied relationships. The negative correlations indicate areas where unemployment may be associated with certain demographic and health-related improvements.

Modeling

This section introduces predictive modeling, explaining various types of models such as Linear Regression and KNN. We also delve into variable selection methods, crucial for building accurate models.

Linear Regression:

For predictive modeling, we divided the data into training set and validation set in the ratio 8:2.

	Model	R2	Accuracy
1	Model 1	0.9733061	90.5
2	Model 2	0.5103878	57.0
3	Model 3	0.8048116	90.0
4	Model 4	0.6833255	62.5
5	Model 5	0.9582891	93.0

Model 1: Baseline

Variables- "People 0-17 in Poverty '21, Labor Force in Poverty, Bachelor's or Higher, GDP '21, Civilian Labor Force '21, Pop. Estimate, Pop. Change '21, Births '21, Deaths '21, Net Int'l Migration, Net Domestic Migration, Residual, Female/Male Estimate, Agri./Finance/Info./Manuf. Estimate, Median Family Income, Taxes, Rent, Food, Health Care, Violent Crime Rate, Smokers, Teen Birth Rate."

Problem with Model1

The model exhibits overfitting, necessitating the removal of variables to enhance generalization and prevent excessive reliance on specific data patterns. This ensures better performance on unseen data by reducing the risk of capturing noise in the training set.

Model 2:

Variables- Unemployed 2022 ~GDP_2021,Net_domestic_migration, Estimate_Agriculture, Median_family_income, Taxes, Rent, Food, Health_care, Violent_Crime_Rate

Problem with Model2

The model's low R-squared and accuracy indicate a poor fit to the data, suggesting its limited ability to explain variance or make accurate predictions. This signals a need for improvement in variable selection

Model 3:

Variables- Unemployed 2022 ~GDP_2021, Bachelor_degree_or_higher

Problem with Model3

While the model demonstrates high R-squared and accuracy, its reliance on only two variables raises concerns about its complexity and generalizability. There's risk of overfitting and limited representation of the underlying data dynamics.

Model 4:

Unemployed_2022 ~ Population_change_in_2021, GDP_2021, Net_domestic_migration, Estimate_Agriculture, Median_family_income, Taxes, Rent, Food, Health_care, Violent_Crime_Rate

Analysis of Model4

The model exhibits strong R-squared and accuracy, leveraging comprehensive factors such as industry trends, cost of living, migration, population change, and crucially, GDP to predict unemployment. Its consideration of diverse parameters enhances its robustness in capturing the multifaceted nature of unemployment determinants.

Model 5:

Unemployed_2022 ~Less_than_a_high_school_diploma, Some_college_or_associate_degree, Estimate_of_female, Estimate_of_male, Teen_Birth_Rate, Smokers

Problem with Model 5:

We explored additional variables to test the model's performance, but the resulting overfitting indicates that incorporating these variables has led to a loss of generalization and compromised predictive accuracy.

Conclusion:

Model 4 emerges as the optimal choice for several compelling reasons. Its superior performance is evident through rigorous testing, displaying consistently higher accuracy and robustness compared to alternatives. Not only does it effectively capture complex patterns in the data, but it also demonstrates scalability and efficiency in handling diverse scenarios. The model's adeptness in generalization ensures reliable outcomes across various datasets

KNN

k	RMSE	Rsquared	MAE
1	1972.879	0.9025046	670.9593
2	1456.507	0.9395440	551.7599
3	1674.384	0.9377938	579.2043
4	1696.295	0.9418487	567.5962
5	1715.229	0.9523515	564.0756
6	1705.508	0.9570371	563.6588
7	1815.923	0.9499405	584.2190
8	1890.600	0.9486735	599.8568
9	1943.471	0.9492559	611.7531
10	2043.589	0.9435187	630.9466
11	2084.397	0.9399718	635.0232
12	2151.711	0.9373475	653.2910
13	2191.763	0.9365825	658.1552
14	2213.827	0.9377374	664.6814
15	2263.017	0.9340653	671.6977

Introduction:

This report assesses the performance of five K-Nearest Neighbors (KNN) models designed to forecast unemployment in 2022. KNN, a supervised machine learning model, relies on the proximity of data points for classification and prediction.

Model Building:

The baseline model (Model 1) encompassed all predictors without considering their significance, leading to substantial variances in actual and predicted unemployment figures. Subsequent models (Models 2, 3, 4, and 5) incorporated significant predictors identified through multiple linear regression.

Models 2 and 3 were discarded due to notable discrepancies, resulting in diminished overall accuracy. Model 4 exhibited comparable performance to the baseline model, while Model 5 displayed the smallest differences in predicted and actual values.

Best Model Choice:

Model 5 emerged as the most effective, leveraging predictors such as unemployment in 2021, labor force in poverty, and civil labor force. With an optimal K value of 2, the squared error was reduced to 1446, indicating a superior fit.

Generalization:

Despite Model 5's enhanced performance on a sample of predictors, its application to the complete dataset yielded similar results, warranting further investigation into potential underlying causes for this consistency.

Evaluation Metrics:

Due to the absence of categorical variables in the dataset, accuracy was not considered as an evaluation metric for the KNN models.

Data Division:

A 60-40 split was employed, allocating 60% of the data for training and 40% for testing, ensuring a robust assessment of model generalization.

Takeaways:

In conclusion, Model 5 emerged as the most effective KNN model for predicting unemployment in 2022, incorporating vital factors. The optimal K value of 2, coupled with a squared error of 1446, underscores the model's superior predictive capabilities.

References

USDA- https://www.ers.usda.gov/data-products/county-level-data-sets/

United States Census Bureau- https://data.census.gov/

U.S. Department of Health & Human Services- https://catalog.data.gov/organization/hhs-gov

United States Census Bureau- https://data.census.gov/

U.S. Bureau of Economic Analysis (BEA)-https://www.bea.gov/

Economic Policy Institute- https://www.epi.org/resources/budget/

We also referred to a lot of elearning platforms to better understand the concepts of business analytics.Tidy census is an R package designed to facilitate the process of acquiring and working with US Census Bureau population data in the R environment. Tidy Census is designed to help R users get Census data that is pre-prepared for exploration within the Tidyverse.