Predictive model with a comparative analysis of the correlative effects socio-economic factors play in relation to life expectancy

Task 3

by

Evan Boyd

Bachelors of Applied Science : Data Management & Analytics Capstone Task 3

Submitted to School of Technology Department

Western Governors University

July 2024

A. Proposal Overview

Research Question and Findings

The primary objective of this capstone project was to investigate the relationship between corruption and life expectancy or *"How does corruption affect life expectancy"* Contrary to initial expectations, the analysis revealed that corruption has a minimal impact on life expectancy when compared to other factors such as geographic region, disease prevalence, sanitation, and carbon dioxide emissions. In fact, corruption was ranked as the second least influential variable in the predictive model.

Methodology and Data

To explore this relationship, a data-driven approach was employed using publicly available World Bank data. The project encompassed standard data science methodologies including cleaning, preprocessing, exploratory data analysis, feature selection, model development, and evaluation. The focus was on identifying correlations and building a predictive model, rather than establishing causal relationships.

Model Development and Evaluation

A random forest regression model was selected for its ability to handle complex relationships within the data. Visualization techniques such as heatmaps, bar charts, and tables were used to explore patterns and communicate findings effectively. The model demonstrated strong performance, with an R-squared value of 0.987 indicating a

high degree of variance explained. However, it's important to note that while the z-score and p-value suggest statistical significance, further validation is required to assess the model's practical implications.

Limitations and Future Work

This project represents an initial exploration of the relationship between corruption and life expectancy. Due to time constraints, the scope was limited to data analysis and model development. Future work could involve deploying the model in a production environment, integrating additional data sources, and conducting further analysis to establish causal relationships.

B. Project Plan

The project initially adhered to a waterfall methodology, with distinct phases for data acquisition, cleaning, exploration, modeling, evaluation, and reporting. This structured approach enabled efficient completion of each phase within the allocated timelines. Data preprocessing and exploratory analysis were finalized within the first week, followed by model development and tuning over the next two weeks. The final phase, encompassing model evaluation and report generation, was successfully concluded within the designated time frame. However, subsequent feedback necessitated a shift from the waterfall to an agile methodology. This adaptive approach allowed for iterative

improvements based on ongoing evaluation. While extending the overall project timeline, it ultimately enhanced the project's quality and robustness. Key factors contributing to the project's success include effective time management, rigorous data handling, and a methodical approach to model evaluation. Challenges such as data quality issues and model overfitting were mitigated through data imputation, outlier treatment, feature engineering, and regularization techniques. Overall, the project demonstrated the value of a structured approach while highlighting the importance of adaptability in response to evolving requirements.

Methodology

C. Data Selection and Collection Process

The data selection and collection phase adhered closely to the project plan, with the World Bank dataset, accessed through Kaggle, serving as the primary data source. This alignment facilitated a smooth transition from the planning to the execution phase. While the World Bank dataset provided a comprehensive overview of global socioeconomic indicators, several challenges were encountered. Data inconsistencies, including missing values and outliers, required meticulous cleaning and preprocessing. Additionally, a notable absence of corruption data in regions such as Sub-Saharan Africa and the Middle East and North Africa limited the dataset's scope. To address these

issues, data extraction and preparation were conducted using Python and Pandas. This combination offered the necessary tools to efficiently handle large datasets and perform complex data manipulations. Techniques such as data imputation, outlier detection, and feature engineering were employed to enhance data quality and create informative variables.

D. Data Extraction and Preparation Process

The data extraction process involved retrieving the World Bank dataset from Kaggle and importing it into a Python environment using Pandas. This facilitated subsequent data manipulation and analysis. Key steps in the data preparation process included:

- **Data Cleaning:** Addressing missing values through imputation methods (e.g., mean, median, mode imputation) and handling outliers using statistical techniques (e.g., z-score, interquartile range).
- **Data Transformation:** Creating new variables, scaling features, and handling categorical data through encoding or one-hot encoding.
- **Data Validation:** Ensuring data consistency, accuracy, and completeness through verification checks and statistical summaries.

Python's Pandas library provided a robust framework for these tasks, offering efficient data manipulation and analysis capabilities.

E. Data Analysis Process

Analytical Methods

The data analysis process comprised descriptive statistics, exploratory data analysis (EDA), and inferential statistics. Descriptive statistics were used to summarize data characteristics, while EDA techniques, including visualizations and correlation analysis, helped uncover patterns and relationships. Inferential statistics, such as hypothesis testing and regression analysis, were employed to draw conclusions about the population based on the sample data.

Tools and Techniques

Python-based libraries, including NumPy, Pandas, Matplotlib, Seaborn, and Scikit-learn, provided the essential tools for data analysis. These libraries offered a comprehensive suite of functionalities for data manipulation, visualization, and statistical modeling. The choice of these tools was driven by their efficiency, flexibility, and wide adoption within the data science community. These tools provided an advantage in processing large amounts of raw tabular data. NumPy, Pandas, Matplotlib, and Seaborn made editing(pre-processing) and visualizing my findings simple. While Scikit-learn provided me with the machine learning capabilities necessary for developing a predictive model. These tools are limited in that they aren't able to implement the model within a program or webpage. Though the implementation of the model in a production environment is outside of the current project scope, so the limitations did not affect the project outcome.

Application of Analytical Methods

Predictive and descriptive analytics were the core methodologies employed in this study. Descriptive analytics, using techniques such as summary statistics and data visualization, provided a foundational understanding of the dataset. Key metrics, distributions, and correlations were explored to identify potential patterns and relationships. Descriptive analytics were used throughout the project, namely during the first 3 objectives of “Acquire data from source”, “Create visualization charts”, “Clean and Preprocess data “. The charts themselves, as well as descriptions of each charts findings, and the conclusion at the end of the .ipynb file all serve as products of my use of the descriptive analytical method.

Building on this foundation, predictive analytics was applied to forecast life expectancy based on the identified variables. Executing the 4th objective, “Build a model that accurately predicts life expectancy based on socio-economic factors”, required the use of predictive analysis. Regression analysis, specifically, was utilized to model the relationship between corruption and life expectancy, considering other relevant factors. The model's performance was evaluated using metrics such as R-squared, mean squared error, and root mean squared error.

Results

F. Evaluation of Project Success

The model's predictive power, as indicated by the high R-squared value of .987, suggests a strong fit to the data. However, the statistical significance of the corruption variable is questionable. A z-test result of -1.325 and corresponding p-value of .185 indicate that the observed relationship between corruption and life expectancy might be due to chance. While the model demonstrates strong predictive performance, its practical implications are limited by the uncertain significance of the corruption variable. Other factors, such as regional disparities, disease prevalence, and sanitation, appear to have a more substantial impact on life expectancy. The model's ability to identify other significant factors influencing life expectancy provides valuable insights for policymakers and researchers. Practically, a young person who'd like to live a long life might use these findings to influence where they plan on moving; playing different scenarios out with the model to project how a country might fair in the future, or even determine where their own country's life expectancy may be headed if changes occurred to observed features like corruption, healthcare expenditure, or CO2 output. The project successfully developed a predictive model with strong explanatory power, and the pckl file developed should please shareholders. However, the lack of conclusive evidence regarding the causal impact of corruption on life expectancy necessitates further investigation.

G. Summary of Key Takeaways

The analysis revealed a complex relationship between corruption and life expectancy. While the model effectively captures underlying patterns in the data, the causal impact of corruption remains inconclusive. Factors such as region, disease prevalence, and sanitation emerged as more significant determinants of life expectancy.The use of visualizations, including heatmaps, bar charts, and tables, was instrumental in communicating complex findings.Combined with the descriptions throughout the .ipynb file, these charts and tables powerfully added to the story hidden within the dataset and without them the project would not relay meaning as easily. These visual aids effectively conveyed the relationship between variables and supported the identification of key trends.To enhance the understanding of the relationship between corruption and life expectancy, future research should focus on:

- **Strengthening causal inference:** Employing advanced statistical techniques and experimental designs to establish causal relationships.
- **Expanding the dataset:** Incorporating additional variables, such as governance indicators and healthcare spending, to improve model performance and explanatory power.
- **Conducting in-depth regional analyses:** Investigating regional disparities in the relationship between corruption and life expectancy to identify specific contexts where corruption has a more pronounced impact.

By addressing these recommendations, future research can provide a more comprehensive and nuanced understanding of the factors influencing life expectancy.

Project Sources

Amuka, J. I., Asogwa, F., Omeje, A., Onyechi, T., & Ugwuanyi, R. (2018). Climate change and Life Expectancy in a Developing Country: Evidence from Greenhouse Gas (CO2) Emission in Nigeria . International Journal of Economics and Financial Issues, 8(4).

Chavan, S. S. (2023a, September 5). Life expectancy & socio-economic (world bank). Kaggle. https://www.kaggle.com/datasets/mjshri23/life-expectancy-and-socio-economic-world-bank

Dreher, Axel and Herzfeld, Thomas, (June 2005). The Economic Costs of Corruption: A Survey and New Evidence. Available at SSRN: https://ssrn.com/abstract=734184 or http://dx.doi.org/10.2139/ssrn.734184

Indicators. World Bank Open Data. (2020). https://data.worldbank.org/indicator

Uchendu, F. N., & Abolarin, T. O. (2015). Corrupt practices negatively influenced food security and live expectancy in developing countries. The Pan African medical journal, 20, 110. https://doi.org/10.11604/pamj.2015.20.110.5311