**Relationship between Economic Characteristics and Telephone Scam Activities—Data Analysis and Predictive Modeling Research**

Fangfang Liu

Saint Louis University

AA 5963: Applied Analytics Master’s Project - III

Dr. Srikanth Mudigonda

July 14, 2024

**Introduction**

In the current digital era, telephone scams have rapidly become a widespread and increasingly rampant crime, causing significant economic and psychological harm to victims. As technology advances, scammers continuously explore new tactics and strategies to increase their success rates, including targeting regions with specific economic characteristics. The adoption of these strategies reveals a more complex situation—different regional economic conditions may influence the frequency of scam calls, the themes used, and the economic impact.

This study focuses on three primary objectives: analyzing the relationship between economic characteristics and the frequency of scam calls, exploring the association between economic characteristics and scam strategies, and assessing the potential impact of regional economic conditions on the economic losses caused by telephone scams. The methods employed include systematically collecting and preprocessing data, performing descriptive statistical analysis, and conducting predictive modeling.

**Literature Review**

The behavioral patterns of telephone scams are noteworthy because, despite the variety of scam activities, they often share similar themes and methods. According to a report by the Federal Trade Commission (FTC), the five most common telephone scam tactics include non-existent prizes, impersonating government agencies, business and investment fraud, and payment methods that are difficult to recover (Fair, 2023). Additionally, the timing of these scams may be meticulously planned. Telecom scammers often choose to call victims in high-expenditure areas after working hours, with noticeable patterns in call timing (Kilinc, 2021).

Furthermore, there are significant regional differences in the reported frequency of telephone scams across the United States. The FTC's "2022 Consumer Sentinel Network Data Book" highlights substantial variations in the complaint rates per 100,000 population among different states. For example, the reporting rate in Georgia is approximately three times higher than in North Dakota (2023). This indicates that, despite the diversity in scam methods, data analysis can still identify scam strategies tailored to different regions. Moreover, research suggests a negative correlation between the level of economic development and scam activities (Omidi, 2017).

Therefore, while telephone scam behavioral patterns are varied, it is still possible to summarize their characteristics through data analysis. However, in the literature I reviewed, I did not find targeted studies on the relationship between telephone scams and regional economic differences.

**Research Methodology**

The data sources for this study include economic data from the U.S. Bureau of Economic Analysis (BEA), phone scam data from the U.S. Federal Trade Commission (FTC), and educational level data from the National Center for Education Statistics.

Since the research question involves two parts, descriptive statistical analysis and modeling, the FTC's dynamically updated scam data is used in the descriptive statistical analysis. This data, which is reported daily to the FTC, provides detailed insights into phone scam strategies, such as the timing of calls, the use of automated voices, and the themes of scams. Since the website only retains datasets for 26 working days, I decided to use the most recent month's data set for a stable analysis. I will download and merge the scam data set from April 24 to May 24. Economic data for this period is based on the BEA's 2023 economic data for each state.

For the modeling part, I will build models based on previously verified correlations between scam themes and regional economics. Although the dataset of real-time phone scam information from the FTC contains detailed information, it is limited in size. Therefore, for modeling, we will use the FTC's annual summary of phone scam information for each state in 2022. To match the year needed for modeling, we will use the 2022 economic data published by the BEA and supplement it with the 2022 educational level data for each state released by the National Center for Education Statistics (NCES).

* Data Collection and Preparation

The data sources for this study include economic data from the U.S. Bureau of Economic Analysis (BEA) and telephone fraud data from the U.S. Federal Trade Commission (FTC). The process of data collection, cleaning and preparation is described below:

1. Telephone fraud data:

Data collection: Collect data on all telephone fraud reports within a one-month period.

Data cleansing: remove duplicate entries and irrelevant columns, retaining only key variables

Dealing with Missing Values: Remove rows with missing values to ensure data integrity.

Time format conversion: convert `Violation\_Date` columns to date-time format for time period analysis.

2. Economic data:

Data collection: obtain state economic data published by the BEA.

Data Cleaning: Remove extraneous columns and focus on key economic indicators such as per capita personal income.

Data Transformation: Organize data by state through pivot tables and update column names for clarity and consistency.

Preprocessing: Remove special characters from the data and rename columns to ensure data consistency.

* Data Conversion and Merging

Due to the differences in format and structure between the BEA economic data and the FTC telephone fraud data, the datasets were merged by sharing variables (states). The data were merged separately for targeted analysis based on the research questions.

1. Call time analysis:

Categorize phone fraud records by time period (morning, afternoon, evening, night) based on the `Violation\_Date` column.

The frequency and percentage of calls for each time period are calculated and merged with economic data.

2. Robot voice usage analysis:

Filter telephone fraud data for the use of recorded or machine voice.

Calculate the percentage of these calls in each state and merge with economic data.

3. Fraud theme analysis:

Categorize telephone fraud by theme (e.g., government impersonation, technical support).

Percentages of each fraud type in each state are calculated and merged with economic data.

* Statistical analysis

A variety of statistical analyses were conducted to explore the relationship between economic characteristics and telephone fraud activity:

1. Descriptive statistics and data visualization:

Descriptive statistical analysis is the foundation of data analysis. It provides a basic description of the data and helps people gain an initial understanding of the dataset's fundamental characteristics (Heiberger & Holland, 2015). Calculating basic descriptive statistics of the dataset helps to understand the distribution and central tendencies of the variables. Additionally, data visualization can help provide a more intuitive view of the basic conditions of the data.

2. Correlation analysis:

By calculating the Pearson correlation coefficient, we assess the strength and direction of the relationship between economic indicators and telephone scam activities. This method reveals the linear relationships between variables, helping to determine the direction for subsequent modeling (Pal & Bharati, 2019). The effectiveness of correlation analysis lies in its ability to intuitively display the connections between variables, providing a theoretical basis for building predictive models.

* Model building and evaluation

To further quantify the relationship between economic variables and telephone scam activities, various modeling techniques are used. By employing advanced modeling techniques such as Random Forests and Lasso Regression, it is possible to handle complex data and predict future trends (James et al., 2023).

1. Multiple predictive models:

Multiple predictive models such as Random Forest, Lasso, Multiple Output Regression, Gradient Boosting, and Neural Networks are used to analyze and predict telephone fraud activity.

2. Model optimization and comparison:

In this study, I will evaluate the performance of various models based on the accuracy metrics (R-squared, MSE, RMSE, and MAE), runtime, and memory usage of their training and testing sets. By comparing the accuracy, efficiency, and stability of different models, the optimal model will be selected (Kuhn & Johnson, 2013). This step ensures the reliability and practicality of the research results by optimizing model performance through identifying overfitting and other issues. The most suitable model will be chosen based on its predictive capabilities and practical application needs.

* Findings and conclusions

Some key insights were gained through data analysis and modeling:

1. Robotic voice use:

The correlation between per capita personal income and the proportion of recorded or machine-dialed calls used in telephone fraud was extremely weak and insignificant.

2. Call time analysis:

The correlation between telephone fraud activity and per capita personal income across time periods was also extremely weak and insignificant.

3. Fraud theme:

There is a significant correlation between different types of fraud and per capita income. For example, energy, solar, and utility types of fraud are significantly positively correlated with per capita income, while medical and prescription types of fraud are significantly negatively correlated with per capita income.

These findings suggest that there are significant differences in the distribution of different types of telephone fraud across states, but the direct correlation with economic data still needs to be further verified.

In summary, although a number of correlations between economic characteristics and telephone fraud activity were identified, these relationships are complex and require further in-depth analysis to validate the correlations and understand the mechanisms behind them.

**Analysis and Discussion of Results**

* Results of statistical analysis

This study reveals some key findings by analyzing economic data with telephone fraud data. The results are analyzed and discussed below for the different research questions.

Firstly, in descriptive statistical analysis, there are noticeable differences in the number of scam calls received across different regions, and the frequency of use varies among scam themes. Secondly, the analysis involves calculating the correlation between personal income and different scam strategies. In the correlation test, the focus is mainly on two test values: the Pearson correlation coefficient and the P-value. The Pearson correlation coefficient is used to assess the linear relationship between two variables. Its value ranges from -1 to +1, where +1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no linear relationship (Nicewander & Rodgers, 1988). The effectiveness of correlation analysis lies in its ability to quantify the strength and direction of the linear relationship between variables. Additionally, significance testing (such as the P-value) is used to determine whether statistical results are statistically significant, with a P-value less than 0.05 generally considered significant (Hatcher, 2013). This test helps determine whether the observed effects could merely be due to chance.

1. Analysis of machine voice usage

In exploring the relationship between per capita personal income and the proportion of recorded or machine dialed numbers used in telephone fraud, I found that:

Correlation analysis: The Pearson correlation coefficient tends to 0 with a p-value of 0.899. this indicates that the correlation between per capita personal income and the proportion of recorded or machine dialed numbers used in telephone fraud is very weak and insignificant.

Linear regression analysis: The R-squared value is close to zero, further confirming that personal income per capita has little effect on the proportion of recorded or machine dialed numbers used in telephone fraud.

Conclusion: These results suggest that personal income per capita is not a major factor influencing the use of recorded or machine dialed numbers in telephone fraud. This finding suggests that other, more influential factors need to be considered when developing preventive measures.

2. Analysis of call duration

In analyzing the relationship between telephone fraud activity and per capita personal income over different time periods, I found that:

Correlation Analysis:

The correlation coefficient for the afternoon hours is 0.026 (p-value = 0.854).

The correlation coefficient for the evening hours was 0.209 (p-value = 0.141).

The correlation coefficient for the morning hour was -0.132 (p-value = 0.356).

The correlation coefficient for the night hours was -0.009 (p-value = 0.948).

Conclusion: These results indicate that the correlation between telephone fraud activity and per capita personal income is weak and insignificant for all time periods. Economic characteristics do not appear to significantly affect the timing of telephone fraud, implying that the temporal distribution of telephone fraud may depend more on other factors, such as the fraudster's behavioral patterns or the victim's work schedule.

3. Fraud theme analysis

In exploring the relationship between different types of telephone fraud and per capita personal income, I found that:

Significant positive correlation:

Energy, Solar & Utilities type of fraud (correlation coefficient = 0.499, p-value < 0.001).

Home improvement and cleaning type of fraud (correlation coefficient = 0.491, p-value < 0.001).

Significant negative correlation:

Medical and prescription type fraud (correlation coefficient = -0.527, p-value <0.001).

Other types of fraud: For example, the correlation coefficient for the type of telephone fraud that impersonates a government, business, or friend or relative is 0.272 (p-value ≈ 0.054), which is positive but does not reach the level of significance.

Conclusion: These results suggest that the distribution of different types of telephone fraud varies significantly across states and that some types of fraudulent activity are significantly correlated with per capita personal income. For example, energy and home improvement types of phone fraud are more common in areas with higher per capita incomes, while healthcare fraud is more frequent in areas with lower incomes. This suggests the need to differentiate strategies for different types of fraud when preventing telephone fraud.

* Build Model

In this study, I build and evaluate a variety of predictive models, including multiple linear regression, Lasso regression, random forest (RF), gradient boosting (GB), and neural network models. Next I will describe the reasons for choosing to build each model:

1. Multiple linear regression

Multiple linear regression aims to predict the outcome of a dependent variable as a linear combination of multiple independent variables (James et al., 2023). When the relationship between economic data features and scam activities may exhibit a linear distribution, multiple linear regression can provide a direct explanation of causal relationships, making it suitable for preliminary exploration of relationships among data.

2. Lasso regression model

The Lasso regression model incorporates a regularization term (the sum of absolute values) in the estimation process, which penalizes the coefficients, thereby enabling variable selection and controlling complexity to help reduce model overfitting (Tibshirani, 1996). When faced with a large number of potentially correlated economic indicators, Lasso regression helps filter out the most influential features through regularization, reducing overfitting and enhancing the interpretability and generalization capability of the model.

3. Random Forest model

The Random Forest model is an ensemble technique that includes multiple decision trees, enhancing the model's generalization ability by introducing randomness during the training process. Each decision tree is trained on a random subset of the dataset, and the final output is the average prediction result of all trees (Breiman, 2001). Given that phone scam data may involve complex classification problems and nonlinear features, Random Forest improves prediction accuracy by constructing multiple decision trees, making it well-suited for addressing randomness and irregularities in classification and regression tasks.

4. Gradient Boosting model

The Gradient Boosting model is a technique that iteratively trains decision trees to minimize the loss function. In each iteration, a new tree is constructed based on reducing the residuals of all previous trees (Friedman, 2001). Since data on phone scams and economics often contain outliers or nonlinear relationships, Gradient Boosting improves prediction accuracy by progressively optimizing the loss function, making it particularly effective in modeling complex data relationships.

5. Neural Networks model

Neural networks are extensively parallel interconnected networks composed of adaptive simple units that attempt to recognize patterns and reason by mimicking the behavior of biological neural systems (Goodfellow et al., 2017). Neural networks typically consist of multiple layers (input layer, hidden layers, and output layer), each containing multiple neurons. For identifying potentially complex patterns and relationships, such as hidden indicators in scam calls or analyzing voice and text data, neural networks can provide deep insights.

* Modeling Result

Since the runtime and memory usage of each model are minimal, the main performance evaluation of the models will rely on accuracy metrics such as R-squared, MSE, RMSE, and MAE. The comparison standards for these model accuracy metrics are as follows:

1. R-squared

R-squared measures the proportion of variance in the observed data that is explained by the model. Mathematically, it is defined as the ratio of the regression sum of squares to the total sum of squares (James et al., 2023). An R-squared value closer to 1 indicates that the model has a strong explanatory power and better predictive performance. A lower R-squared value suggests that the model has not captured the variability in the data effectively.

2. MSE

Mean Squared Error (MSE) is the average of the squares of the deviations between predicted values and actual values in the data. It is one of the standard metrics for evaluating model performance and is often used as a loss function during model parameter optimization (Wackerly et al., 2008). A smaller MSE indicates that the discrepancy between the model's predictions and the actual values is smaller, suggesting better predictive performance of the model.

3. RMSE

Root Mean Squared Error (RMSE) is the square root of the Mean Squared Error (MSE). It provides an error estimate in the same units as the original data, making it easier to interpret and understand (Hyndman & Koehler, 2006). Similar to MSE, a smaller RMSE indicates better model performance because it signifies smaller prediction errors.

4. MAE

Mean Absolute Error (MAE) measures the average of the absolute differences between predicted values and actual values. Compared to MSE, MAE is less sensitive to outliers (Willmott & Matsuura, 2005). A smaller MAE indicates that the average deviation between the model's predictions and the actual values is smaller, suggesting better predictive performance.

| Model | MSE Train | MSE Test | R-Squared Train | R-Squared Test |
| --- | --- | --- | --- | --- |
| Linear Model | 3799680.28 | 10279735.7 | 0.85 | 0.69 |
| Best Lasso Model | 6792162.99 | 6813601.01 | 0.78 | 0.77 |
| Best RF Model | 2193164.35 | 10743368.63 | 0.94 | 0.77 |
| GB Model | 5395.3 | 25167487.34 | 1 | 0.59 |
| Best Neural Network Model | 29133984.54 | 38760396.13 | 0.46 | 0.54 |

| Model | RMSE Train | RMSE Test | MAE Train | MAE Test |
| --- | --- | --- | --- | --- |
| Linear Model | 1949.28 | 3206.20 | 1052.13 | 1504.04 |
| Best Lasso Model | 2606.18 | 2610.29 | 1180.93 | 1282.75 |
| Best RF Model | 1480.93 | 3277.71 | 536.52 | 1626.70 |
| GB Model | 73.45 | 5016.72 | 41.73 | 2194.9 |
| Best Neural Network Model | 5397.59 | 6225.78 | 1939.32 | 2697.93 |

By comprehensively comparing the accuracy metrics between the training and testing sets across different models, we can assess the performance and generalization ability of each model. Below is the ranking of the models along with a discussion:

1. Best Lasso Model: The Lasso model performs best in terms of stability and generalization ability, with close performance on the training and test sets, showing strong prediction of unknown data.

2. Best RF Model: The random forest model has excellent performance on the training set, but the performance degradation on the test set indicates an overfitting problem.

3. Linear Model: The performance of the linear regression model declined on the test set, but still showed a moderate fit.

4. GB Model: The gradient boosting model performs almost perfectly on the training set, but the performance drops dramatically on the test set, showing a strong overfitting problem.

5. Best Neural Network Model: The neural network model performs relatively poorly among all models, both on the training and test sets.

Conclusion: Combining the performance of the models, the Lasso regression model was found to perform best in predicting telephone fraudulent activities with high stability and generalization. While the Random Forest and Linear Regression models also performed well in some aspects, the overfitting problem and adaptive differences require further optimization and improvement. The gradient boosting and neural network models did not perform as well in this study, but still have potential in dealing with larger and more complex data. Future research can further optimize the parameter and feature selection based on the existing models to improve the predictive ability and usefulness of the models.

**Conclusion**

This study draws the following key conclusions from an in-depth analysis of the relationship between economic characteristics and telephone fraud activity: the findings indicate that there is a significant positive correlation between total income and population size and telephone fraud activity, and that telephone fraud activity is more frequent in areas with higher incomes and larger populations; per capita income is relatively weakly related to telephone fraud activity and the correlation is not significant across time; while per capita income affects telephone fraud activity to some extent, it plays a relatively small role; there are significant differences in the distribution of different types of telephone fraud activity across states involving energy, solar and utilities, and home improvements and improvements. The relationship between per capita income and telephone fraud activity is relatively weak and the correlation is not significant across time, and while per capita income affects telephone fraud activity to some extent, it plays a relatively small role; the distribution of different types of telephone fraud activity varies significantly by state, with types of fraud related to energy, solar and utilities, and home improvement and cleaning showing a significant positive correlation with per capita income and types of fraud related to medical care and prescription types showing a negative correlation. These findings suggest that telephone fraudsters may prefer more economically developed and densely populated areas for their fraudulent activities, and that preventive measures need to be tailored to the local context, with targeted prevention strategies for areas with different economic characteristics.

**Limitations and Directions for Future Work**

* Limitations

Since the results of this study in revealing the relationship between economic characteristics and telephone fraud activity, a number of limitations remain. First, the data in this study only cover telephone fraud activity over a specific time period and in a specific region, and may not be fully representative of the nation as a whole or on a larger scale. In addition, the study focused primarily on the impact of economic characteristics (e.g., per capita income, total income, and population size) on telephone fraud activity, but telephone fraud activity is also influenced by other factors. Finally, telephone fraud patterns and tactics evolve over time and with technology. This study is based on historical data and may not accurately predict future fraud trends. These factors were not adequately considered in this study, which may lead to limitations in the results.

* Directions for Future Work

Future research could build on the existing results. First, the scope of the dataset should be expanded to include different time periods and a wider range of regions to enhance the generalizability and representativeness of the results. Second, future studies should incorporate data from more dimensions, such as society, culture, technology level and policy, to provide a more comprehensive analysis. Finally, future research needs to track and update data on an ongoing basis to improve the accuracy of predictions and to address dynamically changing fraud patterns.

**Reference**

Breiman, L. (2001, October). Random forests | machine language. https://dl.acm.org/doi/10.1023/A:1010933404324

Fair, L. (2023, February 23). *FTC crunches the 2022 numbers. see where scammers continue to crunch consumers.*Federal Trade Commission. https://www.ftc.gov/business-guidance/blog/2023/02/ftc-crunches-2022-numbers-see-where-scammers-continue-crunch-consumers

Federal Trade Commission. (2023, February). *Consumer Sentinel Network Data Book 2022*. Federal Trade Commission. https://www.ftc.gov/reports/consumer-sentinel-network-data-book-2022

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep learning*. The MIT Press.

Hatcher, L. (2013). *Advanced statistics in research: Reading, understanding, and writing up data analysis results*. ShadowFinch Media, LLC.

Heiberger, R. M., & Holland, B. (2015). Statistical Analysis and Data Display: An Intermediate Course with Examples in R. https://link.springer.com/book/10.1007/978-1-4939-2122-5

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. https://doi.org/10.1016/j.ijforecast.2006.03.001

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023, June 30). *An introduction to statistical learning*. SpringerLink. https://link.springer.com/book/10.1007/978-1-0716-1418-1

Kilinc, H. H. (2021). A case study on fraudulent user behaviors in the telecommunication network. *Electrica*, *21*(1), 74–84. https://doi.org/10.5152/electrica.2021.20050

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. SpringerLink. https://link.springer.com/book/10.1007/978-1-4614-6849-3

Nicewander, A., & Rodgers, J. L. (1988, February). Thirteen ways to look at the correlation coefficient. https://www.stat.berkeley.edu/~rabbee/correlation.pdf

Omidi, M., Min, Q., & Omidi, M. (2017). Combined effect of economic variables on fraud, a survey of developing countries. *Economics &amp; Sociology*, *10*(2), 267–278. https://doi.org/10.14254/2071-789x.2017/10-2/20

Pal, M., & Bharati, P. (2019, July 19). *Introduction to correlation and linear regression analysis*. SpringerLink. https://link.springer.com/chapter/10.1007/978-981-13-9314-3\_1

Regression shrinkage and selection via the lasso. (n.d.). https://webdoc.agsci.colostate.edu/koontz/arec-econ535/papers/Tibshirani%20(JRSS-B%201996).pdf

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. http://www.jstor.org/stable/2346178

Tu, H., Doupé, A., Zhao, Z., & Ahn, G.-J. (2019, August 16). *Users really do answer telephone scams*. USENIX. https://www.usenix.org/conference/usenixsecurity19/presentation/tu

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical statistics with applications*. Thomson Brooks/Cole ; Cengage.

Willmott, C. J., & Matsuura, K. (2005, December 19). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. https://www.jstor.org/stable/24869236

Xing, J., Yu, M., Wang, S., Zhang, Y., & Ding, Y. (2020, August 28). *Automated fraudulent phone call recognition through deep learning*. Wireless Communications and Mobile Computing. https://www.hindawi.com/journals/wcmc/2020/8853468/

Yonder. (2023, March 16). Online scams and fraud research: Summary report. https://www.ofcom.org.uk/\_\_data/assets/pdf\_file/0025/255409/online-scams-and-fraud-summary-report.pdf