**Comprehensive Analysis of Construction Spending Using Machine Learning: An XG Boost Approach**

**1. Introduction**

The construction industry plays an important role in country’s economic growth, GDP and development projects are highly impacted by spending decisions and patterns. Use of predictive analysis allows designated stakeholders to anticipate spending trends and make informed decisions and base their decisions on data driven insights . This comprehensive report provides an extensive data analysis and predictive spending on designated months using XG Boost.In addition to predictive accuracy this report consists of insights crucial for gaining trusts for future spendings in this sector.

The report is divided into 9 sections , introduction, problem statement, literature review,dataset description, data cleaning, exploratory data analysis, feature engineering, preprocessing steps before applying XG Boost, implication of XG Boost to forecast future spendings. All 9 sections are covered comprehensively with code snippets attached with sections involving coding.

**2. Problem Statement**

There are mainly 2 purposes of this analysis is. First, it aims to uncover patterns and trends in historical construction spending data, providing a clearer picture of how spending has evolved over time(represented by various nature of graphs such as boxplot , scatter plot and heatmaps). Second, it seeks to build a reliable predictive model capable of estimating future spending levels, using insights drawn from past data.

By achieving these objectives, the analysis will offer actionable insights into the dynamics of construction spending.

**3. Literature Review**

Predictive modeling is an important tool specially in construction industry as stakeholders tend to do long term planning with predictions , time series analysis was used before but it fell short working with complex datasets .

Machine learning algorithms had emerged and made predictions on complex datasets feasible and for example (XG Boost); it does not only rely on historical trends instead it also considers other different factors such as weather , labor availability, material costs to get a deeper understanding of data.

It’s a step forward for construction sector , where forecasting can contribute to cost savings and resource allocation. The ability to handle real-world complexities and deliver fast, reliable predictions makes XGBoost a preferred choice for this dataset

1. **Dataset Description**

The dataset contains 169 rows and 119 columns, providing data about construction spending over time. Here's an overview of its structure:

**Key Features:**

1. Time Information:

* time.index: Index of the time period.
* time.month: Numeric representation of the month.
* time.month name: Name of the month (e.g., "Jan").
* time.period: Combination of month and year (e.g., "Jan2002").
* time.year: Year (e.g., 2002).

1. Spending Categories:

* Columns such as annual.combined.amusement and recreation, annual.combined.commercial, etc., represent annual combined spending in various categories like amusement, commercial, communication, etc.
* Detailed breakdowns of spending for categories like educational, conservation, and development are also provided.

1. Public and Private Spending:

* current.public.nonresidential, current.public.office, etc., detail public sector spending in specific areas.
* Spending areas include power, public safety, religious, residential, and infrastructure projects like sewage and waste disposal, transportation, and water supply.

1. Spending Totals:

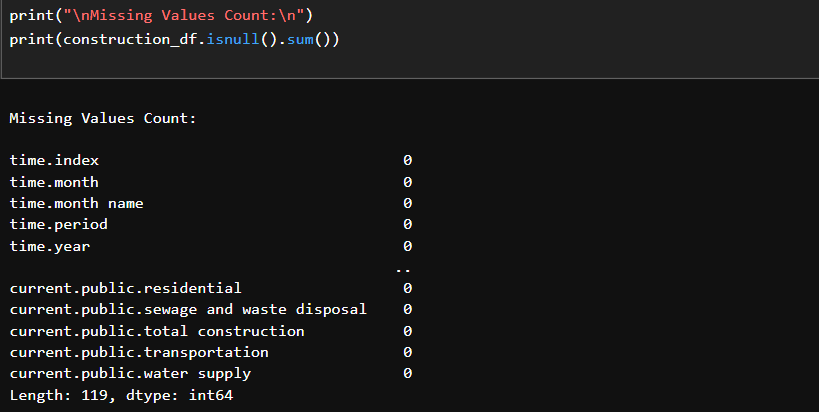
* Columns such as current.public.total construction provide aggregated totals for specific categories.

A screenshot of a black screen

Description automatically generated

**5) DATA CLEANING**

There were no missing values in the data



Data had around 119 columns , in which a large number of them didn’t contribute to any significant data insights , and a significant number had single constant values , those columns were dropped during data cleaning

**6) EXPOLATORY DATA ANALYSIS (EDA)**

**Trend of Total Construction Spending Over Time (2002–2016)**

The above plot illustrates the trend in total construction spending over time, with data points covering multiple years. The x-axis represents the time period, while the y-axis shows the spending levels.

**Key Observations:**

1. **Seasonal Patterns:** The spending curve shows a clear seasonal pattern, with periodic peaks and troughs occurring annually. This indicates a recurring trend likely influenced by factors such as weather, budget cycles, or economic conditions.
2. **Overall Growth:** Over the years, there is a general upward trend in construction spending, reflecting economic expansion or increased infrastructure investment.

**Insights:**

* The observed seasonality in construction spending is critical for predictive modeling, as it suggests that temporal features like month and year are highly influential.
* The fluctuations during specific periods highlight the need for external economic factors to be incorporated into the model for better accuracy.
* Understanding these trends helps policymakers and stakeholders align construction activities with optimal time frames to leverage peak periods of investment.

**Total Construction Spending by Public Sector**

**Description:** The bar chart provides a clear representation of how much money is spent on construction across different public sectors. Each bar stands for a particular sector, with its height indicating the total amount allocated to that sector. This visualization makes it easier to compare spending between sectors at a glance, highlighting which areas receive more investment and which receive less

**Insights:**

* **Highway and Street Investments**: The significant focus on highway and street spending underscores the priority placed on infrastructure development. This trend likely reflects the growing demand for improved transportation networks and the ongoing expansion of urban areas. Investments in these sectors are essential for supporting economic growth, reducing traffic congestion, and connecting communities more effectively.
* **Public Sector Prioritization:** The spending patterns suggest that education and health care are critical focus areas, aligning with government priorities for societal welfare.
* **Opportunity for Diversification:** The minimal spending in certain sectors, such as recreation and public office, could indicate potential areas for future investment to balance sectoral development.

**Monthly Variations in Public Construction Spending**

**Description:**

The boxplot above tells how public construction spending varies throughout the year. Along the x-axis, are the months, while the y-axis displays the corresponding spending amounts. This visualization captures median spending for each month, the interquartile range (IQR), and any outliers that stand out.

**Key Observations:**

* Throughout the year, spending shows a consistent upward trend, reaching its highest levels between June and October. However, in winter, particularly in November, December, and January, there’s a noticeable drop in spending..

**Variability Across Months:**

* During the winter months, such as January and February, spending tends to be more stable, as seen in the narrower interquartile ranges (IQRs). This suggests that spending levels are more predictable during this time. On the other hand, the summer months display more variability. The wider boxplots and whiskers indicate that there’s greater fluctuation in spending.

**Insights**

**Strategic Planning:**

* Understanding this seasonality is essential for stakeholders to allocate resources efficiently and schedule construction activities during peak months for optimal productivity.

**Correlation Heatmap of Construction Spending Features**

**Description:**

The heatmap above gives an overview of the correlations between different features in the dataset. The color intensity indicates how strongly and in which direction the features are related:

* Dark red shows a strong positive correlation (close to 1).
* Dark blue indicates a strong negative correlation (close to -1).
* Lighter colors suggest a weaker or no correlation (close to 0).

By examining this heatmap, you can easily see how various features are connected to each other, as well as their relationship with the target variable, which in this case is the total construction spending.

**Key Observations**

**Strong Positive Correlations:**  
Some features, like current.public.highway and current.public.total construction, show a strong positive correlation, which makes sense given that highway spending is a significant part of overall public construction expenditure. Similar patterns emerge between related categories, such as annual.combined.highway and current.public.highway, further highlighting the close relationship between these spending areas.

**Weak Correlations:**  
Certain features, including annual.private.lodging and other private-sector spending categories, have weaker correlations with public construction spending. This suggests that the private sector’s influence on public spending is relatively limited.

**Negative Correlations:**  
While there are a few features that show negative correlations, these are generally weak, meaning that any inverse relationships are minimal and not very impactful.

**Insights:**

* **Temporal Features:** Time-based variables, like month and time.index, exhibit varying correlations with sector-specific spending, suggesting that trends and seasonality play a role in how spending patterns unfold over time.
* **Policy and Decision-Making:** Policy makers can recognize sectors with strong correlations, that is, Project A is closely related to Project B can guide them in prioritizing investments.

**7) Feature Engineering**

Feature engineering is selecting, manipulating and transforming raw data so it can significantly improve the accuracy of machine learning models. In this section, we will walk through the techniques used to create new features from the dataset.

**1. Time-Based Features**

Extracting year, month, quarter, dayofyear and week from *time.period* column, we have provided the model with detailed and key time-related features that help it identify recurring patterns and seasonal changes.

**2. Lagged Features**

Lagged features provide historical context, enabling the model to account for momentum and dependencies in spending. A lagged feature is created by taking the value of a variable at a previous time point and including it as a feature in the model at the current time point:

 These lag features capture spending from the last 1–3 time periods, which are crucial for short-term forecasting for investment.

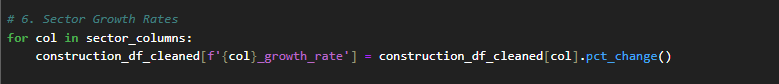
**3. Rolling Averages**

Using rolling averages helps reduce short-term fluctuations, allowing the underlying trends to become clearer over time by calculating series of averages of different selections of data from full data set.

 **3-period and 6-period rolling averages** helps reduce noise in data and emphasize medium-term patterns

**4. Sector Growth Rates**

Sector-specific growth rates capture temporal variations within each sector:



 These features provide clear insights into sectoral changes over time helping the organization make decisions depending on the currently well performing sector.

**5. Seasonality Indicators**

Seasonal indicators encode cyclical patterns:

 Binary flags for winter and summer help the model adjust for seasonal variations in construction activities, help create future plans with considering the factor of seasonal variations.

**6. Cyclic Features**

Cyclic transformations ensure smooth transitions between the beginning and end of cyclic periods:

 Sine and cosine transformations are used to capture the cyclical pattern of months, converting this information into a format that the model can understand and process effectively.

**8) Preprocessing Steps Before Applying XGBoost**

Preprocessing steps such as mapping, cyclic transformations, and encoding are essential to ensure that the dataset is in an optimal format for machine learning models like XGBoost.

**1. Mapping Numeric Months to Names and Back**

* Converting months back to numeric form prepares them for cyclic feature engineering.

**2. Cyclic Transformation of Months**

* It is been observed that months follow a cyclical pattern, where December is closely linked to January. Sine and cosine transformations help capture this cyclical relationship, allowing the model to treat the months as part of a continuous loop when they are closely related.  
  Without these transformations, representing months numerically (e.g., 1–12) could mislead the model into thinking there's a linear progression, which may distort how the features relate to each other.

**3. One-Hot Encoding for Non-Numeric Features**

* Machine learning models such as XGBoost require numeric data for learning, so one-hot encoding is used to transform categorical variables into binary columns, making them compatible with the model. For Example, the gender column from Male and Female is transformed into 0 and 1.

**4. Aligning Training and Test Sets**

* In order to ensure the model works smoothly, it is important that both training and test datasets have the same columns after encoding or feature engineering. Misalignment can lead to prediction errors, as the model expects the data in a consistent format.

**9) Implementation of XGBoost for Predicting Construction Spending**

Description:

This code provides a full workflow for preparing the data and applying the XGBoost model to predict the current.public.total construction values. The process covers checking the target column, managing missing data, splitting the dataset for training and testing, scaling features, and training the XGBoost regressor. The outcome is a DataFrame showing the predicted construction spending values for each period in the test set.

**Predicted Construction Spending Results**

The table above shows the predicted construction spending values for each month from January 2015 to January 2016, based on the XGBoost model. The predictions show a clear trend of rising spending during the middle months of the year, with peaks occurring around mid-year (May through August) and a decline toward the winter months.

**Predicted vs Actual Construction Spending Comparison**

The graph above compares the predicted construction spending values against the actual values over the same period. The actual values, represented by blue points, closely align with the predicted values, shown as orange points, indicating that the model was successful in capturing the seasonal trends and fluctuations in the data. The comparison reveals that the model effectively captured the key patterns in construction spending, including the peak spending periods during mid-year and the decrease during winter months.

**10) References**

* Yan, H., He, Z., Gao, C., Xie, M., Sheng, H., & Chen, H. (2022). Investment estimation of prefabricated concrete buildings based on XGBoost machine learning algorithm. Advanced Engineering Informatics, 54, 101789.
* Pham, T. Q. D., Le-Hong, T., & Tran, X. V. (2023). Efficient estimation and optimization of building costs using machine learning. International Journal of Construction Management, 23(5), 909-921.
* Ghimire, P., Pokharel, S., Kim, K., & Barutha, P. (2023, June). Machine learning-based prediction models for budget forecast in capital construction. In Proceedings of the 2nd International Conference on Construction, Energy, Environment & Sustainability, Funchal, Portugal (pp. 27-30).
* ShashwatWork. (n.d.). *Construction spending dataset*. Kaggle. Retrieved November 20, 2024, from <https://www.kaggle.com/datasets/shashwatwork/construction-spending-dataset>