Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/BA08.20250230
Published: December 20, 2025

Prediction of Smartphone Sales Revenue through Machine Learning Approach

Salman Monir
Bachelor of Science in Industrial and Production Engineering Khulna University of Engineering & Technology (KUET), Bangladesh
salmanmonir53237@gmail.com

Md. Rafsan Jamil
Bachelor of Science in Industrial and Production Engineering Military Institute of Science and Technology (MIST) rafsanrahat9@gmail.com

Shezan Ahmed
Bachelor of Science in Industrial and Production Engineering Khulna University of Engineering & Technology (KUET), Bangladesh
shezan.kuetipe@gmail.com

Sagor Hossain
Bachelor of Science in Industrial and Production Engineering Khulna University of Engineering & Technology (KUET), Bangladesh
sagorhossain3716@gmail.com

Abstract

Predicting revenue is a crucial concern for defining organizational strategic goals. Business strategies are largely dependent on the status of the product's revenue. Various statistical and analytical techniques are available to forecast business revenue which may not provide a precise prediction due to considering all relevant predictive indicators or features. This study addresses the critical importance of accurately predicting mobile revenue in a highly dynamic and complex market influenced by diverse factors such as product type, regional 5G coverage, and market dynamics. The primary objective was to create an effective forecasting platform that uses powerful machine learning models to predict the nonlinear nature of relationships that drive mobile revenue generation. The novelty in this work is in the design and implementation of an ensemble strategy called Optimal Weighted Blending (OWB) which combines the best Gradient Boosting Machines (GBMs) such as CatBoost, XGBoost, and LightGBM and gets high predictive accuracy and stability. The study utilized an extensive dataset on a quarter of mobile sales of between 2019 and 2024 that has detailed features such as units sold, market share, and features indicators of the 5G ecosystem to perform rigorous preprocessing and feature engineering to prepare the data and train and evaluate the model. It was found that OWB ensemble was better than individual models with an R^2 of 0.91 and a mean absolute error (MAE) of 1.67 million successfully modeled complex interactions among features and mitigated prediction variance. Consequently, the findings of this study would be beneficial to organizations in estimating the revenue of any product, as well as in surviving in a highly competitive world.

Keywords

Machine Learning, Smartphone, Data Analysis, Revenue Prediction, Artificial Neural Network.

© IEOM Society International

1. Introduction

Today's business world is very competitive and are highly dependent on decision-making. Making right decisions at the right time is therefore the most important task for any company and it becomes more crucial for electronic products. Compared to non-electronic products, sales of electronic products vary more. For example, introducing a new smartphone may not always grab the market. In this competitive business world, forecasting revenue is very essential to survive and shines up the business by making the right decision. So, decision-making is essential there [Armstrong and Grohman, 1972]. In other words, forecasting determines what will happen in the future by looking at what has already occurred and what is currently occurring [Butler and Hansen, 1991]. The forecast has so many linear and nonlinear forecasting models available. It has been shown that nonlinear forecasting models have more accuracy or less error rate than linear models. So nonlinear model will help to forecast with great accuracy [Chung et al., 2022] [Gallant, 1975]. A crucial part of many human endeavors now includes social networking through Facebook, Twitter, Insta gram, Tumblr, and other social media platforms. Persistent, archived can be recovered and analyzed the resulting social data. Social data analytics is influencing existing practices in news media, policy, marketing, finance, product development, and entertainment in addition to educating people about them [Malinvaud, 1970]. If some social media can be defined as a second life, then the smartphone has grown into an extension of the human body and mind. Apple, iPhone, and Samsung are the best-selling devices in history and are connected to vast volumes of big data on most social media platforms. Besides, revenue cannot just have a linear relationship. It may also be a nonlinear function [Norstro¨m, 1996]. Proper forecasting can improve product sales, storage management, and business decision performance. reenue forecasting is an in-depth analysis of past performance to help understand how much the business might be bright at that time [Sreemathy and Prasath, 2022]. So, the prediction of revenue with the linear forecast models contains a large error. To overcome this problem a non-linear model to forecast the revenue could be adopted. Thus, machine learning can be exploited for a more reliable prediction of revenue.

Thus, revenue prediction gives a big help for a company to grow up well and sound. Therefore, the objectives of this research are: firstly, to find out the best-performing ML model for predicting smartphone sales revenue; and secondly, to explore the impact of a particular feature on the model accuracy. The rest of the paper is organized as follows: section II discusses the literature review related to this research, section III discusses the research methodologies, section IV designs the model formulation and analysis of results, and section V concludes the research highlights the main outcomes, research limitations and potential scopes for the future research.

2. Literature Review

([Petros¸anu, 2022]) highlights that prior e-commerce sales forecasting research has developed advanced machine and deep learning models (e.g., LSTM, XGBoost, CNN, hybrid models) to achieve superior accuracy compared to traditional methods. The key gap identified is the inability of existing literature to consistently provide fine- grained, long-term accurate sales revenue forecasts. The current paper proposes to address this gap by using a Directed Acyclic Graph Neural Network (DAGNN) architecture to generate long-term, product-category-specific daily sales revenue forecasts. ([Sreemathy and P., 2022]) indicates that prior work successfully employs advanced machine learning models for sales forecasting, achieving high predictive accuracy. The achievement of this paper is demonstrating that integrating consumer perception data (e.g., from reviews) significantly improves the forecast accuracy and provides characterization/insights into the factors driving sales. The central gap addressed is the lack of existing models that can simultaneously predict sales and offer scalable, explainable insights into the con- sumer sentiment that causes those predictions, a deficit this paper aims to solve. ([Mansur and K., 2025]) develop a hybrid Convolutional Neural Network (CNN)-Long Short-Term Memory (LSTM) model that integrates both historical sales data and crucial external factors, such as holidays, weather, and salary days.

This approach successfully fills the gap left by previous models that often failed to simultaneously leverage both complex temporal-spatial patterns and real-world contextual variables, providing a more robust and informed forecasting tool for retail decision-making. ([Hwang and Y.-K., 2023]) addressed the significant research gap in new product sales forecasting, where limited or no historical data is available, making traditional methods ineffec- tive. They leveraged data homogeneity by proposing a framework that first used K-means clustering to segment historical products into homogeneous groups with similar sales patterns, and then developed a tailored ensemble forecasting model (combining Random Forest, XGBoost, etc.) for each specific cluster. Their main achieve- ment was demonstrating superior predictive accuracy—achieving an optimal MAPE of 8.3309% on South Korean smartphone sales data—by showing that cluster-specific models significantly outperform single, universally ap- plied models, thus providing a more scalable and adaptable solution for industries with rapid product lifecycles. ([Hwang and Y.-

© IEOM Society International

K., 2023]) addressed the research gap of new product sales forecasting under data scarcity by successfully identifying the Random Forest (RF) model as the optimal predictor among twelve machine learning options. They overcame the limitation of sparse history by enabling the RF model to learn from the collective features and sales patterns of the entire product category, achieving a strong predictive accuracy with a Mean Absolute Percentage Error (MAPE) of 42.6258%. For future work, the authors plan to enhance the model's ro- bustness by incorporating competitive sales data, consumer sentiment, and exploring deep learning architectures like RNNs to validate its application across other short-lifecycle product categories. ([Zamzami, 2023]) addresses the significant challenge of accurately forecasting the rate of 5G technology adoption, a crucial task for policy-makers and telecom operators. The study fills a gap left by traditional statistical methods, which often struggle with the complex, non-linear relationships inherent in technology adoption data. Zamzami's approach involves utilizing Deep Learning (DL) models, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, to analyze various factors influencing 5G uptake. The key achievement is demonstrating that these DL models offer superior predictive accuracy compared to conventional techniques, providing a more robust and timelier forecast to guide strategic investments and infrastructure planning for the ongoing global 5G rollout. The existing literature on smartphone sales forecasting is generally limited by its narrow use of features, typically including only price, reviews, or sales numbers, with only one prior study expanding this set but still failing to predict revenue based on brand name or brand model. Crucially, the impact of input features has not been explored in depth. Therefore, the current study is specifically designed to fill these gaps by predicting smartphone sales revenue using a robust dataset of six key features and, importantly, providing a focused exploration of the impact of those input features on the final prediction.

3. Methodology

3.1 Data Acquisition

The data employed in this study is an extensive sample of the mobile revenue figures, and it contains numerous parameters connected with the mobile market dynamics, including units sold, market share, 5G coverage, regional data, and subscriber data and it provides quarterly mobile sales performance since 2019 Q1 to 2024 Q4. This data gives a rich pool of features that can be used to predict mobile revenue, particularly the consideration of such variables as market share, regional 5G coverage, and 5G technology adoption. The sources of data will include publicly available records and industry reports which will make it relevant and timely to the study.

3.2 Data Preprocessing

- **Cardinality and Frequency Analysis:** Analysis of the frequency and count of unique records ensuring whether there is a balanced number of records in the in various product lines and geographical locations.
- **Missing Values:** We addressed the missing or inconsistent values by detecting them in the course of the EDA. The analysis did not consider any of the rows that contained no target variables.
- **Distributional and Outlier Analysis:** Visualizing the distribution of all the numerical variables through histograms, Kernel Density Estimates(KDE) and box plots to determine skewness and outliers. Key continuous variables (e.g., Units Sold, Revenue ($) were tested using the Shapiro-Wilk test to formally test the deviants of a normal distribution.
- **Correlation Profiling:** Producing scatter plots and a detailed correlation table of all the numerical attributes. This showed major relationships, including the natural correlation between Revenue (in dollars) and Units Sold.
- **Anomaly Detection:** Particular emphasis has been placed upon data anomalies identified in the market intelligence features. These observations bore out the observation of composite or unrealized indexing which required the non-linear models to be used which were resilient to such characteristics of data.

3.3 Feature Engineering and Data Splitting

Given the mixed nature of the dataset, standardized preprocessing steps were essential to prepare the features for machine learning algorithms:

Feature Scaling (Normalization): All the input features were numeric, and Standard Scaling was applied to them with the help of the Standard Scaler transformer. This process put the data around the mean of zero data and normalized the variance of the data to one avoiding features that have larger inherent scales.

Categorical Encoding: Quarter, Product Model, 5G Capability, Region are nominal categorical features, which

© IEOM Society International

were encoded through One-Hot Encoding via the One-hot-Encoder in a Column transformer. This trans- formed the discrete categorical labels into a sparse matrix of binary columns which could be used by the mathe- matical models with no ordinal bias introduced.

3.4 Models Used

A set of Seven different regression models based on diverse machine learning paradigms were chosen as base learners to have diversity and reduce biases in each model:

3.4.1 k-Nearest Neighbors Regressor (KNN)

The k-Nearest Neighbors (KNN) Regressor is a non-parametric model classifier, which uses the mean of the nearest k training samples in the feature space to obtain the target value. A distance measure is normally used to determine the proximity between the samples including Euclidean distance. KNN has no assumptions regarding the underlying data distribution which is helpful in non-linear relationships. Nevertheless, high-dimensional data may lead to a decrease in its performance, as the curse of dimensionality sets in.

3.4.2 Random Forest Regressor

Random Forest is an ensemble learning algorithm that is created to construct a number of decision trees throughout the training and returns the average forecast of all trees in a regression task. It minimizes the possibility of overfit- ting by averaging all trees, each of which were trained on a different portion of the data through random sampling. Random Forest has been noted to have strong strength, adaptability, and capability of processing categorical and numerical data. It also offers the measures of feature importance, which can be useful in the interpretation of the model.

3.4.3 XGBoost Regressor

XGBoost (Extreme Gradient Boosting) is an optimized high performance gradient boosting algorithm. It constructs a set of decision trees in a sequence fashion, in which each tree corrects the mistakes of the preceding decision tree. XGBoost will come up with the concept of regularization, tree pruning, and parallelization to enhance speed and accuracy. It is very popular with structured/tabular data and is characterized by high-performance and scalability and has frequently been used with great success in several machine learning contests.

3.4.4 CatBoost Regressor

CatBoost (Categorical Boosting) is a gradient boosting algorithm that works directly on categorical features, without one-hot encoding or other preprocessing. It employs an effective method to deal with categorical variables to compute their statistics and consider them in the model. CatBoost has also been characterized by the way it can use large datasets that have both numerical and categorical traits and achieve high performance with little preprocessing. It also implements methods of minimizing overfitting and thus is robust in most applications in the real world.

3.4.5 LightGBM Regressor

Another speed and efficiency-optimized gradient-boosting machine is LightGBM (Light Gradient Boosting Machine). It is a large dataset algorithm and offers quick training with high accuracy. LightGBM splits decision trees with a histogram-based approach, thereby minimising the memory and computation time. It can particularly be applied to large datasets and high-dimensional data. LightGBM has also categorical features and is efficient in dealing with missing data as well as being parallelized thus it is a good fit in high-performance applications.

3.4.6 Artificial Neural Network (ANN)

The Artificial Neural Network(ANN) that is used to solve this regression problem is a fully connected (Dense) feed-forward network that involves learning of complicated, non-linear, mappings between the engineered input features and the continuous target variable (Mobile Phone Revenue). The model has three layers with a reduced number of neurons (128, 64, 32) and the Rectified Linear Unit (ReLU) activation function in every hidden layer to bring about non-linearity to prevent the vanishing gradient issue. The shape of the input to the network is given by the ultimate quantity of preprocessed features and the output layer is a single neuron with a linear activation function, typical of regression issues. The Adam optimizer with the learning rate of 0.001 was used to optimize

© IEOM Society International

the given model, and the Mean Absolute Error (MAE) was used as the main loss function, which was being trained during 100 epochs with a batch size of 32.

3.5 Final Ensemble Strategy: Optimal Weighted Blending

The single best-performing model from the baseline evaluation was combined with other top-performing tree- based models to create a definitive forecast using a novel ensemble technique: Optimal Weighted Blending (OWB).

Algorithm for Optimal Weighted Blending (OWB): Optimal Weighted Blending is a mathematically robust algorithm that is aimed at minimizing the aggregate predictive power of an ensemble. Instead of using simple averaging, OWB uses the determination of weights as a constrained maximization.
The objective is to find a set of weights (w_i) that minimizes a chosen error metric (the loss function) for the final combined prediction:

$$\hat{Y}_{blend} = w_1\hat{Y}_{Cat} + w_2\hat{Y}_{XGB} + w_3\hat{Y}_{LGBM}$$

where $\hat{Y}_i$ is the prediction of the i^{th} base model.

1. **Objective Function:** The Mean Absolute Error (MAE) was selected as the minimization target, focusing the ensemble on reducing the average magnitude of absolute prediction errors across the test set.

$$Minimize: MAE(w) = \frac{1}{N}\sum_{j=1}^{N}\left|y_j - \hat{Y}_{blend,j}\right|$$

2. **Constraints:** The optimization was performed under strict conditions to ensure the resulting blend is stable and correctly scaled:

   - ***Unbiasedness Constraint (Equality):*** The sum of all individual model weights must equal 1. This ensures the prediction remains scaled correctly:

   $$w_1 + w_2 + w_3 = 1$$

   - ***Non-Negativity Constraint (Bounds):*** All weights must be non-negative:

   $$0 \le w_i \le 1$$

3. **Optimization Engine:** The optimal weights were derived using the constrained minimization algorithm Sequential Least Squares Programming (SLSQP). This method is suitable for minimizing a multivariate scalar function subject to both bounds and equality constraints.

The weights determined by this optimization yield the lowest possible MAE for a linear combination of the selected base model predictions on the unseen test data.

3.6 Performance Evaluation Metrics

Three standard regression measures were used to measure the final predictive performance of all the base models and the ensemble model:

i. **R-squared (R^2):** The value of R^2 is used to indicate the percentage of variance in the dependent variable that can be explained by the independent variables. Greater values (nearer to 1.0) signify a superior fit.
ii. **Root Mean Square Error (RMSE):** The standard deviation of the prediction errors (residuals). RMSE is very sensitive to outlier since it greatly penalizes large prediction error.
iii. **Mean Absolute Error (MAE):** The average magnitude of the predictions errors of various predictions, disregard- ing their direction. It is an interval score in which all the individual variations receive the same weights that give a measure of average forecast accuracy that can easily be interpreted.

© IEOM Society International

4. Results and Discussion

Data Analysis and Exploratory Data Findings

An in-depth Exploratory Data Analysis (EDA) was implemented to describe the overall complicated nature of the mobile revenue data such as the distributional analysis of all variables and exploration of the linear correlation pattern.

Categorical Feature Analysis and Revenue Distribution

The relationship between key categorical market segments and the target variable, Revenue ($), revealed significant heterogeneity, which strongly influences predictive complexity:

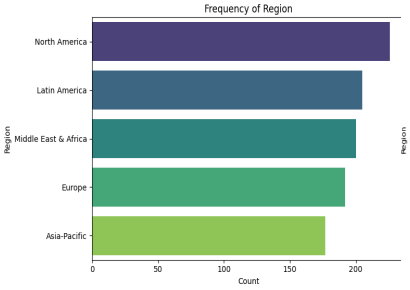

Figure 1. Frequency of Region

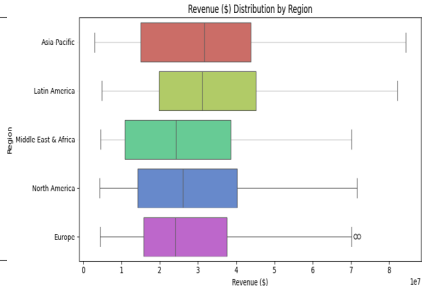

Figure 2. Revenue Distribution by Region

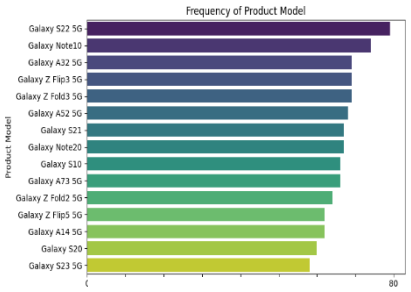

Figure 3. Frequency of Product Model

- Regional Dynamics: The dataset is distributed relatively evenly among the five major territories(Figure 1), with North America having the largest number of observations. However, the revenue distribution by region (Figure 2) shows significant disparities. Europe exhibits the greatest upper quartile range of revenue, while the Middle East & Africa region shows a lower average revenue, possibly due to a lower Average Selling Price (ASP) or a greater volume of low-end products (Figure 3).
- Product Segmentation: Product Model is a high-cardinality feature. Analysis of the revenue distribution (Figure 4) shows that recent foldable devices and certain middle-end A-series products are associated with the largest median revenues and the greatest variability. This validates the need for the predictive model to learn model-specific, high-impact non-linear revenue functions, particularly those related to premium and high-volume 5G product lines.
- Temporal and 5G Influence: Revenue performance by quarter (Figure 5) reveals a high level of volatility of the revenue performance throughout the entire year and high outliers with high revenue are found in all four quarters (Q1, Q2, Q3, and Q4). This is an indication that high-revenue events are not as dictated by calendar season, but rather individual product launch cycles.

Numerical Feature Distributions and Relationship to Target

The analysis of numerical features utilized histograms, box plots, and scatter plots to understand individual distributions and their raw linear relationship with Revenue (Figures 6–13):

Units Sold and Implicit ASP: The Units Sold (Figure 7) is multi-modal with the highest point falling be- tween 30,000 and 40,000 units sold. The scatter plot of the Units Sold and Revenue ($) shows a general, non-deterministic trend. There is a slight positive correlation, but the high revenue peaks (60M-80M) are in the entire range of unit volumes. This is a crucial observation: the forecasting of revenue is not a linear process of the volume, on the contrary, it is very sensitive to the implicitly calculated Average Selling Price (ASP) that needs to be modeled with complex interactions of features (e.g., Product Model, Region, and 5G indicators).

© IEOM Society International

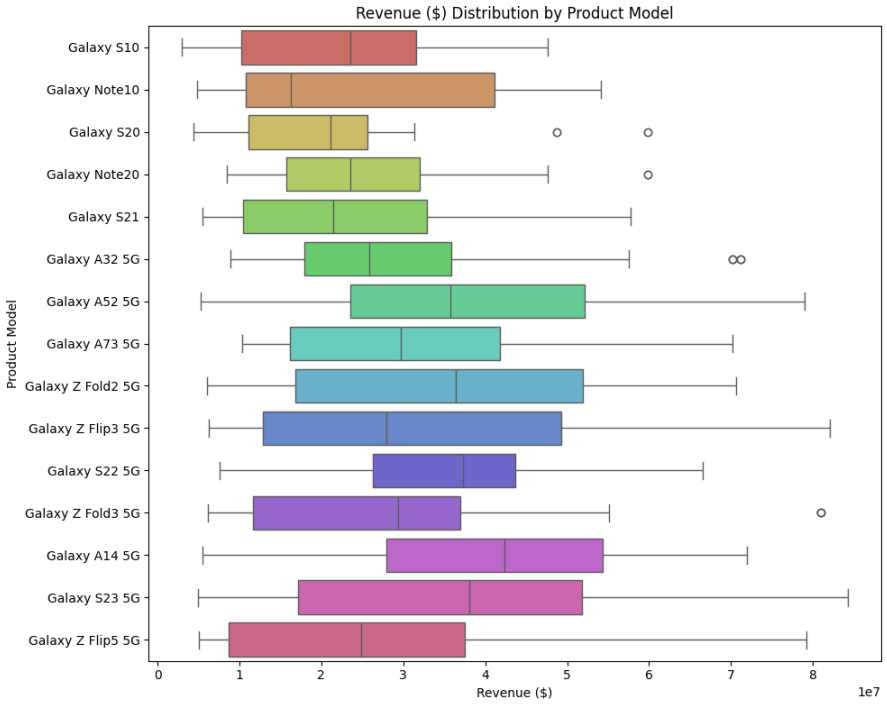

Figure 4. Revenue ($) Distribution by Product Model

• Market Share and Time: Figure 8 Year feature indicates that the number of observations per year is generally evenly distributed through time, however, the scatter plot reveals that there is no simple linear time-series trend which predominates in the data (high revenue data points are distributed across all years, 2019-2024). The Market Share (%) feature (Figure 9) is also skewed to the right and alone there is no distinct linear relationship with the target.

• 5G Ecosystem Indicators: The distributions of the Regional 5G Coverage (%), Avg 5G Speed (Mbps), Preference to 5G (percentage), and 5G Subscribers (millions) (Figure 10, 11, 12, 13) are not normal and wide. It is interesting to note that the max values of Records in Regional 5G Coverage (%) are well over 100% and this

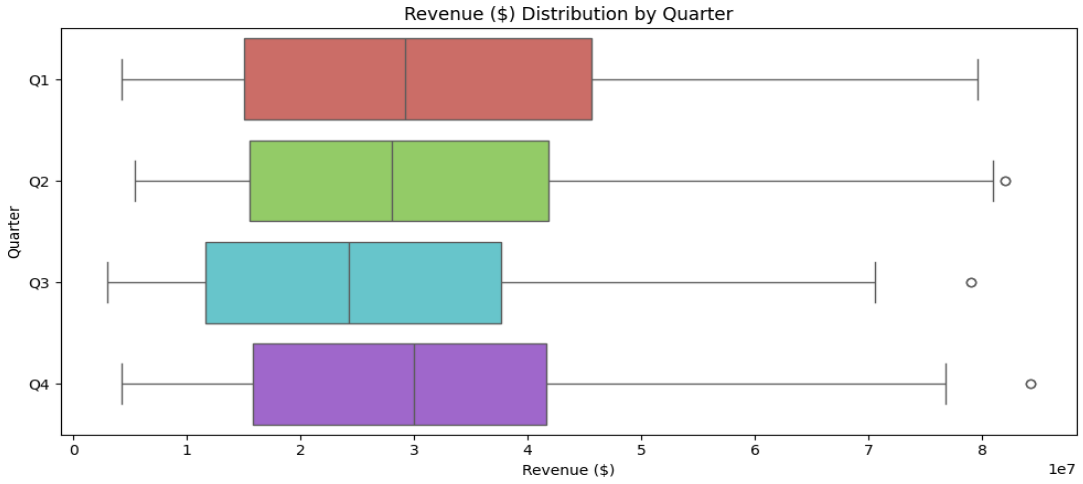

Figure 5. Revenue ($) Distribution by Quarter

© IEOM Society International

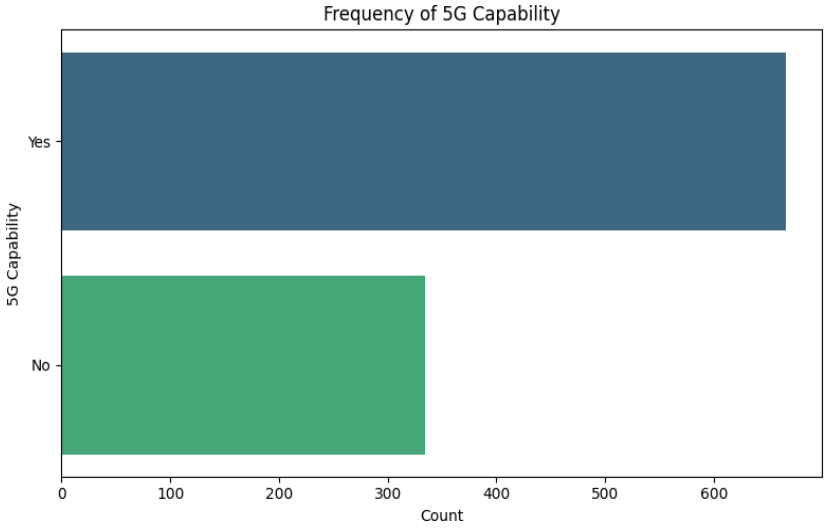

Figure 6. Frequency of 5G Capability

figure confirms that this is a composite measure of market maturity and not just a mere physical coverage measure. Similar to the other numerical characteristics, none of these indicators are strongly, uniquely linearly correlated with Revenue (), although it is possible that they can serve as predictors, as is the case in product and time-related complex interactions.

Correlation Matrix Analysis

The correlation matrix of numerical attributes (Figure 14) formally quantifies the pairwise linear relationships within the dataset, providing objective evidence for the non-linear hypothesis.

Target Correlation: Critically, none of the numerical features exhibit a strong positive linear correlation with Revenue ($) (the right-most column). The highest correlation values recorded are only 0.16 (Preference for 5G (%) and Market Share (%)) and 0.10 (Units Sold). This proves that simple linear models do not reflect the inherent structure of mobile revenue generation, which justifies the strategic change to non-linear tree-based and deep learning models that can learn high-order feature interactions.

1.

Inter-Feature Correlation: The correlation ships between the independent variables show the existence of low collinearity, which is an advantage of the model stability. The strongest inter-feature correlation (0.26) occurs between Market Share (%) and Preference for 5G (%, which is understandable, because they are related variables when it comes to market adoption, but this number is too small to imply the risk of a severe multicollinearity.

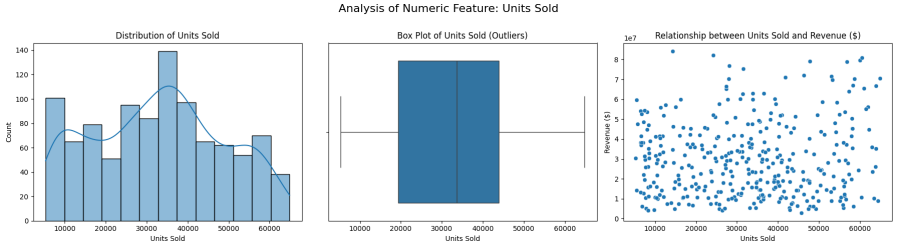

Figure 7. Analysis of Numeric Feature - Unit Sold

© IEOM Society International

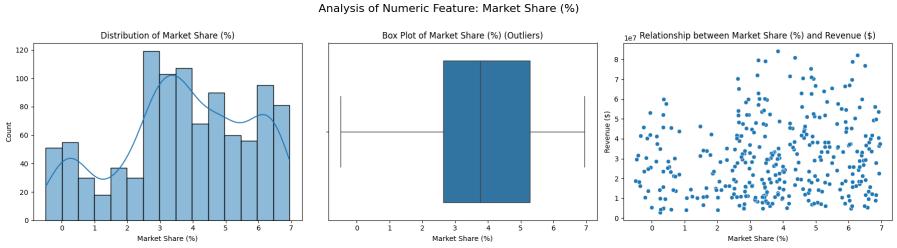

Figure 8. Analysis of Numeric Feature - Market Share(%)

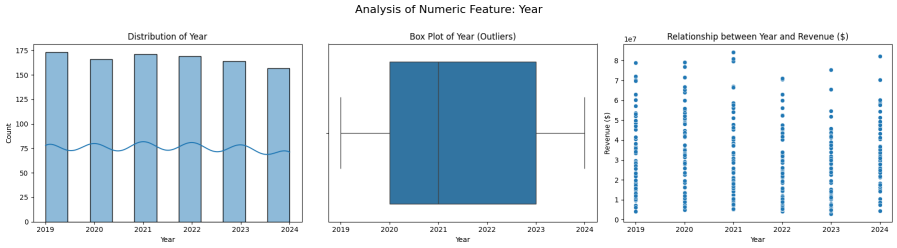

Figure 9. Analysis of Numeric Feature - Year

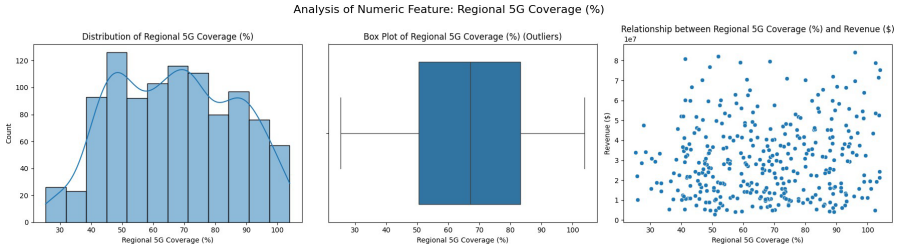

Figure 10. Analysis of Numeric Feature - Regional 5G Coverage(%)

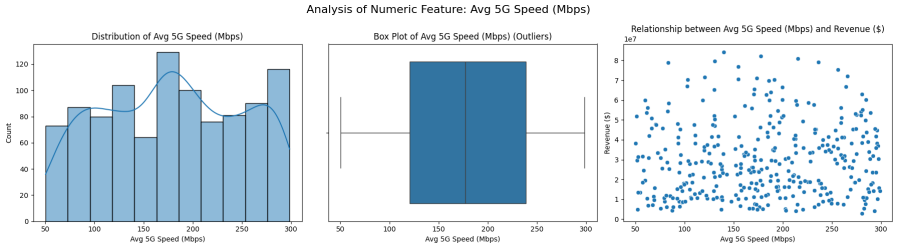

Figure 11. Analysis of Numeric Feature - Avg 5G Speed (Mbps)

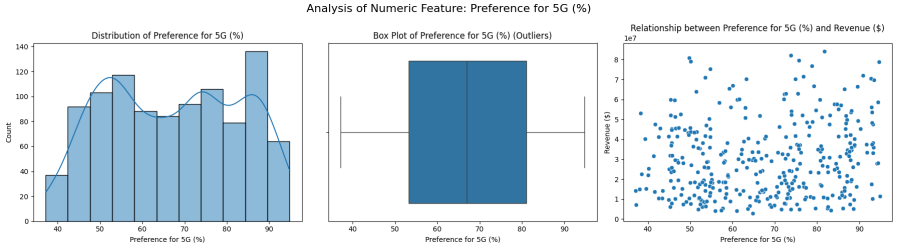

Figure 12. Analysis of Numeric Feature - Preference for 5G(%)

© IEOM Society International

Performance Analysis of Predictive Models

The model assessment stage was meant to methodically benchmark nine single algorithms of regression verifica- tion of the need to find complicated modelling strategies and ended with the creation of a very productive ensemble

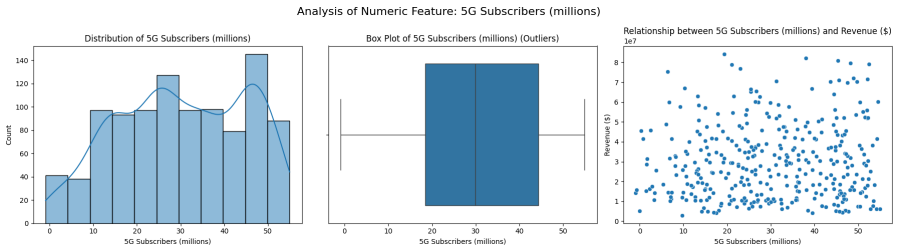

Figure 13. Analysis of Numeric Feature:5G Subscribers (millions)

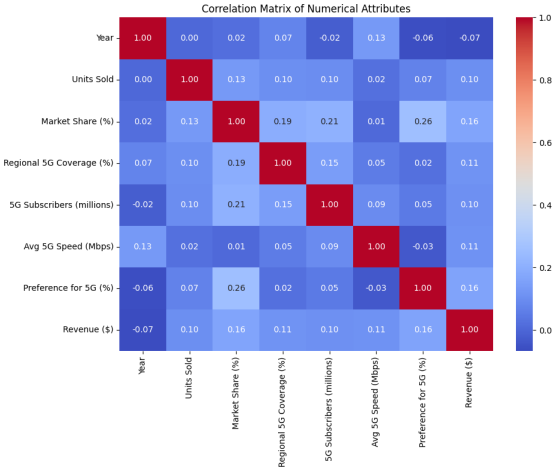

Figure 14. Correlation Matrix of Numerical Attributes

system. The performance measures, which are the R^2 (R Square), Mean Absolute Error (MAE), and the Root Mean Square Error (RMSE), are the quantitative measures of models to explain the non-linear revenue dynamics detected in EDA.

Comparative Analysis of Base Model Performance

The evaluation results established a decisive superiority among the tree-based ensemble methods, which collectively dominated the conventional and instance-based models. The models clustered into distinct performance tiers (Table 1):

Table 1. Model Performance Comparison

Models	R Square	MAE	RMSE
Random Forest Regressor	0.85	4.34E+06	6648485.442
LightGBM Regressor	0.85	3.65E+06	6706372
CatBoost Regressor	0.90	2.66E+06	5461003.841
K-Neighbors Regressor	0.22	1.14E+07	15218399.99
XGBoost Regressor	0.87	1.88E+06	6273649.298
ANN	0.86	3.47E+06	6399854.954
Ensemble Model	0.91	1.67E+06	5267780.356

Gradient Boosting Machine (GBM) Supremacy: The best four models CatBoost, XGBoost, LightGBM and Random

© IEOM Society International

Forest are all essentially non-linear models completing the EDA finding that revenue is a product of complex, and interacting features, and not a product of simple effects. The CatBoost Regressor has the biggest explanatory power with a value of 0.90 and the lowest RMSE of 5.46 million. This high performance is mainly gained through its architecture that is highly optimized to support high-cardinality nominal features, which is Product Model without compromising the information content, thus efficiently representing the implicit variations of the information as the Average Selling Price (ASP) that is inherent in the data.

On the other hand, the XGBoost Regressor has registered the second smallest overall percentage of the prediction error showing an MAE of 1.88 million. This indicator means that, in average, the forecasts made by XGBoost were the most proximate to the actual revenue values. Although CatBoost elucidated a somewhat greater percentage of the total variation (R^2), XGBoost was more dependable in reducing the average prediction error. The other boosting algorithms, LightGBM (R^2: 0.85, MAE: 3.65 million) and Random Forest (R^2: 0.85, MAE: 4.34 million) had strong, high-fidelity predictions, but were not as accurate as the two specialized boosting implementations.

Deep Learning and Model Failure: The ANN model based on Multi-layer perceptron (MLP) architecture also performed competitively with an R^2 of 0.86. The success of this is that the revenue data has rich, global non- linear variations that they can utilize well with the neural network structure to make the ANN a useful, diverse base learner to the GBDT models. Contrarily, the K-Neighbors Regressor did not work at all, the model produced the lowest possible R^2 at 0.22 with a considerably high MAE at 11.40 million. Such a failure emphasizes how simple, instance-based methods cannot provide accurate mapping of the high dimensional and structurally heterogeneous feature space of the mobile market data.

Neighbors Regressor failed entirely, yielding an R^2 of just 0.22 and an exceptionally high MAE of $11.40 million. This failure underscores the inability of simple, instance-based methods to accurately map the high- dimensional and structurally heterogeneous feature space of the mobile market data.

Impact of the Optimal Weighted Blending (OWB) Ensemble: The last step was the synthesis of the predictions of the top three Gradient Boosting Machines (CatBoost, XGBoost, and LightGBM) into one highly stable ensemble using Optimal Weighted Blending (OWB). The motivations behind the use of this technique are to reduce residual and correlated errors among the base models by defining the blending process as a constrained minimization of the final MAE (Table 2).

Table 2. Performance Comparison of Best Models

Models	R-Squared (R^2)	MAE (Million $)	RMSE (Million $)
CatBoost Regressor (Best Base)	0.90	2.66	5.46
Ensemble Model (OWB)	0.91	1.67	5.27

This decrease in the average error (MAE) and sensitivity to large errors (RMSE) is the ultimate mark of the effect of the ensemble. The OWB method managed to detect the decorrelated errors of the three base models and used them effectively to make sure that the resulting blended prediction would be more robust and provide it with better predictive accuracy than any of the individual model architectures. The resulting optimal weights in which XGBoost achieved the greatest proportional weight endorsed that models that have independent error profiles are of paramount importance in reaching the highest stable encompasses regardless of whether they are also the top encompasses as a performance of their own in terms of crude R^2.

5. Conclusions

The paper has created a most accurate and consistent model of mobile revenue prediction achieving a maximum possible R-square of R^2 =0.91 and minimum MAE value. The study presented that mobile revenue is not linear at its fundamental level. Rather, complex variables such as type of product and an implicit Average Selling Price (ASP) affect it, not simply the volume measures. The first benchmarking demonstrated that Gradient Boosting Machines (GBMs) performed better and that CatBoost Regressor had the highest R^2 (0.90) and XGBoost Regressor had the 2nd most accurate predictions (MAE of $1.88 million). The last ensemble technique was Optimal Weighted Blending (OWB), which was a blend of predictions of the best GBMs in a manner that it minimized the sum of the error variance. The OWB model was best performing with a value of R^2 of 0.91 and lowest value of MAE at 1.67 million. This

© IEOM Society International

demonstrates that no other approach is as effective as the ensemble approach when it comes to forecasting stable and predictable revenue in this highly dynamic market. In order to maintain the predictive superiority, the upcoming study should focus on enhancing stability and interpretability. Also, the system should be made stronger by adding exogenous economic indicators to bring in new, unrelated information.

References

Armstrong, J. S. and Grohman, M. C., A comparative study of methods for long-range market forecasting, *Management Science*, vol. 19, no. 2, pp. 211–221, 1972.

Butler, J. E. and Hansen, G. S., Network evolution, entrepreneurial success, and regional development, *Entrepreneurship & Regional Development*, vol. 3, no. 1, pp. 1–16, 1991.

Chung, I. H., Williams, D. W. and Do, M. R., For better or worse? Revenue forecasting with machine learning approaches, *Public Performance & Management Review*, vol. 45, no. 5, pp. 1133–1154, 2022.

Gallant, A. R., Nonlinear regression, *The American Statistician*, vol. 29, no. 2, pp. 73–81, 1975.

Hwang, S. and Go, Y.-K., A sales forecasting model for new-released and short-term product: A case study of mobile phones, *Electronics*, vol. 12, p. 245, 2023.

Malinvaud, E., The consistency of nonlinear regressions, *The Annals of Mathematical Statistics*, vol. 41, no. 3, pp. 956–969, 1970.

Mansur, S. and Khan, S., Sales forecasting for retail stores using hybrid neural networks and sales-affecting variables, *PeerJ Computer Science*, vol. 11, p. e1425, 2025.

Norström, C. J., Break-even analysis with nonlinear revenue functions: A note, *Scandinavian Journal of Management*, vol. 12, no. 2, pp. 159–163, 1996.

Petros, A. D.-M., E-commerce sales revenues forecasting by means of dynamically designing, developing and validating a directed acyclic graph (DAG) network for deep learning, *Electronics*, vol. 11, p. 442, 2022.

Sreemathy, J. and Prasath, N., Machine learning based sales prediction and characterization using consumer perceptions, in *Proc. IEEE International Conference on Smart Information Systems and Technologies (SIST)*, IEEE, 2022.

Sreemathy, J. and Prasath, N., Machine learning based sales prediction and characterization using consumer perceptions, in *Proc. 6th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*, IEEE, pp. 1072–1076, 2022.

Zamzami, I. F., Deep learning models applied to prediction of 5G technology adoption, *Applied Sciences*, vol. 13, p. 4920, 2023.

Biographies

Salman Monir is a graduate student in the Department of Industrial Engineering and Management at Khulna University of Engineering & Technology (KUET), Bangladesh. His research interests include Machine Learning, Operations Research, and Supply Chain Management. He has gained hands-on experience working with advanced techniques such as Large Language Models, Deep Learning Models, and Graph Neural Networks.

Md. Rafsan Jamil is a graduate in Industrial and Production Engineering from the Military Institute of Sci- ence and Technology (MIST), Bangladesh, and has also completed his Master's degree in Applied Statistics and Data Science from Jahangirnagar University. His research interests include Machine Learning, Data Science, Op- erations Research, and Advanced Manufacturing systems. With over four years of professional experience in the manufacturing industry, he has developed a strong foundation in data-driven process improvement and operational efficiency. Beyond academics, Rafsan has been actively involved in impactful co-curricular activities during his undergraduate studies, demonstrating leadership, innovation, and a commitment to continuous growth in both in- dustrial and research environments.

Shezan Ahmed holds a Bachelor of Science in Industrial and Production Engineering from Khulna Univer- sity of Engineering & Technology (KUET), Bangladesh, and is currently pursuing a Postgraduate Diploma in Knitwear Industry Management at BRAC University. His research interests span Supply Chain Management, Ma- chine Learning, Operations Research, Lean Manufacturing, and Sustainability with a focus on Circular Economy principles. Alongside his academic endeavors, Shezan has over 2.5 years of professional experience in leading manufacturing organizations, where he has contributed to improving industrial efficiency, process optimization, and sustainable production practices. He is driven by a strong passion for integrating advanced data-driven solu- tions into industrial systems to foster productivity and environmental responsibility.

© IEOM Society International

Sagor Hossain is a graduate student in the Department of Industrial Engineering and Management at Khulna University of Engineering & Technology (KUET), Bangladesh. His research interests include Machine Learn- ing and Supply Chain Management, with a particular focus on integrating advanced technologies for optimizing supply chains.

© IEOM Society International