Department of Computer Science & Engineering

United International University

Forecasting SMS Traffic and Balance Availability with Machine Learning

Author:

Mohammad Saifur Rahman

ID: 0122420002

Nirupom Das Dipto

ID: 0122230023

Md Mizanur Rahman

ID: 0122420016

Md. Raqibur Rahman

ID: 0122310001

Supervisor:

Dr. Mohammad Nurul Huda

Professor, Head of CSE Dept.

Department of CSE, UIU

Course Details:

Course Code: **CSE 6011**

Course Name: **Data Mining**

Trimester: **Summer-2024**

Program: **MSCSE**

October 2024

Contents

Chapters **Page**

1 Introduction **1**

1.1 Background 1

1.1.1 Objectives 2

1.1.2 The Problem 2

1.1.3 The Solution 2

1.1.4 The Benefits 3

1.1.5 Significance of our project 3

1.2 Literature Review 4

1.3 Research Gap 4

1.4 Objectives 4

2 Methodology **5**

2.1 Corpus Collection 5

2.1.1 Dataset 5

2.1.2 Experimental Setup 5

2.1.3 Description of the data sources used 5

2.1.4 Data Collection Process 6

2.1.5 Data preprocessing techniques 7

2.1.6 Polynomial Regression 7

2.1.7 Key Terms in Polynomial Regression 8

2.1.8 System Diagram 8

2.1.9 Cross-Validation 9

2.1.10 Implementation 9
2.1.11 Cross-Validation Results 9

3 Experimental Results and Analysis **11**
3.1 RESULTS and DISCUSSION 11
3.1.1 Visualization of Predicted vs. Actual SMS Volumes 11
3.1.2 R-squared Analysis 12

4 Conclusion and Future Works **15**
4.1 Conclusion 15
4.2 Future Works 16

Chapter 1

Introduction

This report presents a comprehensive analysis of SMS traffic and balance availability for a specific set of clients, with a particular focus on predicting potential service interruptions during holidays. By utilizing advanced machine learning techniques, such as **polynomial regression**, we aim to accurately forecast future SMS traffic and balance levels, enabling clients to optimize their usage and avoid unexpected service disruptions.

By using advanced data analytics, this study provides a comprehensive forecast of SMS traffic and balance levels, enabling organizations to proactively manage their client communication strategies and resource allocation. This predictive model accurately anticipates SMS traffic during peak periods, such as holidays, and offers hourly updates on balance levels for the next two days, ensuring a seamless and uninterrupted service experience for clients.

Ultimately, this research contributes to the field of machine learning applications and provides practical solutions for organizations seeking to effectively manage their SMS balance allocation to clients, particularly during peak demand periods like holidays.

1.1 Background

At Wintel Limited, there was a need for SMS balance forecasting analysis to predict SMS traffic and balance levels in advance. By anticipating these factors, the company can avoid SMS shortages during peak times, such as holidays or unexpected events, ensuring continuous service availability.

1.1.1 Objectives

- The objectives of this project is to develop a predictive model using advanced machine learning techniques to accurately forecast masking and non-masking SMS traffic and balance levels for a specific set of clients. More specifically the purpose of this project is to generate next 2 days predictions.

- To identify potential service interruptions during peak periods, such as holidays, and provide early warnings to organizations and its clients.

- To enable organizations to optimize their SMS communication strategies and resource allocation by providing timely information on SMS traffic and balance levels.

- To contribute to the field of machine learning applications by demonstrating the effectiveness of polynomial regression in predicting SMS traffic and balance.

- To provide practical solutions for organizations seeking to improve their SMS communication infrastructure and deliver exceptional customer service.

1.1.2 The Problem

In today's fast-paced world, SMS communication has become an integral part of our daily lives. For organizations with a large customer base, managing SMS traffic and ensuring adequate balance availability can be a complex challenge, especially during peak periods such as holidays. Service interruptions due to insufficient balance or overloaded networks can lead to customer dissatisfaction and financial losses.

1.1.3 The Solution

To address these challenges, this report presents a comprehensive analysis of SMS traffic and balance prediction & automation for a specific set of clients. By employing advanced machine learning techniques, such as **polynomial regression**, we aim to develop a predictive model that can accurately forecast future SMS traffic and balance levels. This information will empower organizations to make informed decisions regarding their communication strategies

and resource allocation, ensuring a seamless and uninterrupted service experience for their clients.

1.1.4 The Benefits

- **Improved Service Quality:** Accurate predictions of SMS traffic and balance levels will enable organizations to proactively address potential issues, ensuring a consistent and reliable service experience for their clients.
- **Optimized Resource Allocation:** By understanding future demand, organizations can allocate resources more efficiently, reducing employees hours of struggle during the holidays and minimizing service disruptions.
- **Enhanced Customer Satisfaction:** A reliable and uninterrupted SMS service can significantly improve customer satisfaction and loyalty.
- **Data-Driven Decision Making:** The insights gained from this analysis will provide organizations with a data-driven foundation for making informed decisions about their SMS communication strategies.

This report aims to contribute to the field of machine learning applications and provide practical solutions for organizations seeking to optimize their SMS communication infrastructure and deliver exceptional customer service.

1.1.5 Significance of our project

Forecasting SMS traffic flow using machine learning algorithms offers a powerful method for not only predicting SMS volumes based on historical data but also incorporating various external factors that may affect traffic.

External factors like weather conditions, time of day, and calendar events can introduce irregularities in SMS traffic patterns, making it crucial to account for these variables when building forecasting models.

By integrating these external variables into the prediction model, machine learning algorithms can generate more accurate forecasts. These improved predictions help organizations

make more informed decisions about how to manage SMS traffic flows, optimize resource allocation, and improve operational efficiency.

1.2 Literature Review

1.3 Research Gap

1.4 Objectives

- The objectives of this paper is to develop a predictive model using advanced machine learning techniques to accurately forecast masking and non-masking SMS traffic and balance levels for a specific set of clients. More specifically the purpose of this project is generate next 2 days predictions.
- To identify potential service interruptions during peak periods, such as holidays, and provide early warnings to organizations and it's clients.
- To enable organizations to optimize their SMS communication strategies and resource allocation by providing timely information on SMS traffic and balance levels.
- To contribute to the field of machine learning applications by demonstrating the effectiveness of polynomial regression in predicting SMS traffic and balance.
- To provide practical solutions for organizations seeking to improve their SMS communication infrastructure and deliver exceptional customer service.

Chapter 2

Methodology

2.1 Corpus Collection

2.1.1 Dataset

- **Total Dataset:** 20 incidents (achieved from the work-data till date)
- **Training Data:** 70% of the total dataset
- **Testing Data:** 30% of the total dataset

2.1.2 Experimental Setup

- **Polynomial Degree:** We have considered the best-fit degree of 2 for our model.
- **Accuracy Checking:** We have analyzed whether the real-life SMS traffic requirement lies within the predicted SMS traffic values generated by our model.

2.1.3 Description of the data sources used

For this project, data from Wintel Limited is used. To achieve better results and improve model building, a real-life, real-time dataset will be used, ensuring more accurate and practical outcomes.

Date	Balance
2024-09-24	138706
2024-09-25	138214
2024-09-26	137241
2024-09-27	134192
2024-09-28	118229
2024-09-29	116883
2024-09-30	113839
2024-10-01	112638
2024-10-02	111951
2024-10-03	110807
2024-10-04	105434
2024-10-05	103655
2024-10-06	100337
2024-10-07	96591
2024-10-08	86970
2024-10-09	162602
2024-10-10	155813
2024-10-11	155048
2024-10-12	154230
2024-10-13	154108

Table 2.1: Balance Data from 2024-09-24 to 2024-10-13

2.1.4 Data Collection Process

In wintel each 4 hours a cron/scheduler will be called where this model will work and generate the predictions. For Both masking and non-masking sms traffic and balance seperate models will be called. In each day approximately 6 times the this process will work. We will pick last 180 days dataset from wintext portal mssql database and use it to the model.

2.1.5 Data preprocessing techniques

- Timestamps were converted into a consistent datetime format to facilitate accurate time-series analysis.
- SMS traffic counts and balance levels were ensured to be in numeric formats for proper calculation and analysis.
- Duplicate records in SMS traffic logs were identified by checking for identical timestamps, client IDs, and traffic counts.
- Special care was taken to adjust for traffic and balance patterns during holidays, which often exhibit different behaviors from regular periods. Historical holiday data was cross-referenced with the current dataset to ensure accurate classification and analysis.
- SMS Traffic per Operator & Client: The SMS traffic per operator and client was calculated to analyze usage patterns.
- Balance per Client: The balance per client was calculated to assess balance allocation.
- Balance per Operator: The balance per operator was calculated to assess balance allocation to opertor wise.

2.1.6 Polynomial Regression

Polynomial Regression is an extension of linear regression where the relationship between the independent variable(s) and the dependent variable is modeled as an nth-degree polynomial. This method is particularly effective when the data exhibit non-linear relationships that cannot be captured by a simple linear regression model.
The general form of a polynomial regression equation is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \epsilon$$

Where:

- y is the dependent variable (response variable).

- x is the independent variable (predictor variable).
- $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients of the polynomial.
- ϵ is the error term (residuals).

2.1.7 Key Terms in Polynomial Regression

- **Polynomial Terms**: These are the powers of the independent variable(s), such as x^2, x^3, etc., that enable the model to fit non-linear data.
- **Degree of Polynomial**: This refers to the highest exponent in the polynomial equation (e.g., n in x^n). Higher degrees capture more complexity in the data but can lead to overfitting.
- **Least Squares Method**: The coefficients $\beta_0, \beta_1, \ldots, \beta_n$ are estimated by minimizing the sum of squared residuals between the observed and predicted values.

2.1.8 System Diagram

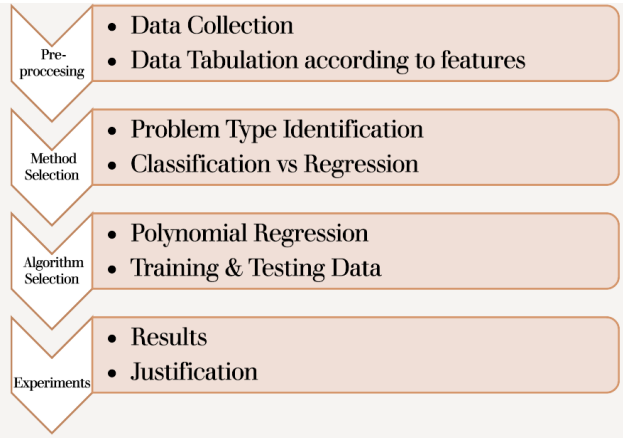

Figure 2.1: System Diagram

2.1.9 Cross-Validation

K-fold cross-validation is a common strategy where the dataset is divided into k subsets (folds). The model is trained k times, each time using a different fold for validation and the remaining folds for training.

2.1.10 Implementation

We use Python's `scikit-learn` library to implement cross-validation. Below is the code demonstrating k-fold cross-validation using polynomial regression:

2.1.11 Cross-Validation Results

To evaluate the performance of the model, we performed a 5-fold cross-validation. The Mean Squared Error (MSE) for each fold is presented below:

Fold	**MSE**
Fold 1	731,548,313.47
Fold 2	51,671,242.87
Fold 3	225,424,809.40
Fold 4	134,718,956.94
Fold 5	377,833,216.14
Average MSE	**304,239,307.76**

Table 2.2: Mean Squared Error (MSE) for Each Fold

The table above summarizes the MSE for each fold of the cross-validation. The average MSE across all folds is approximately 304,239,307.76, indicating the model's performance.

Chapter 3

Experimental Results and Analysis

3.1 RESULTS and DISCUSSION

3.1.1 Visualization of Predicted vs. Actual SMS Volumes

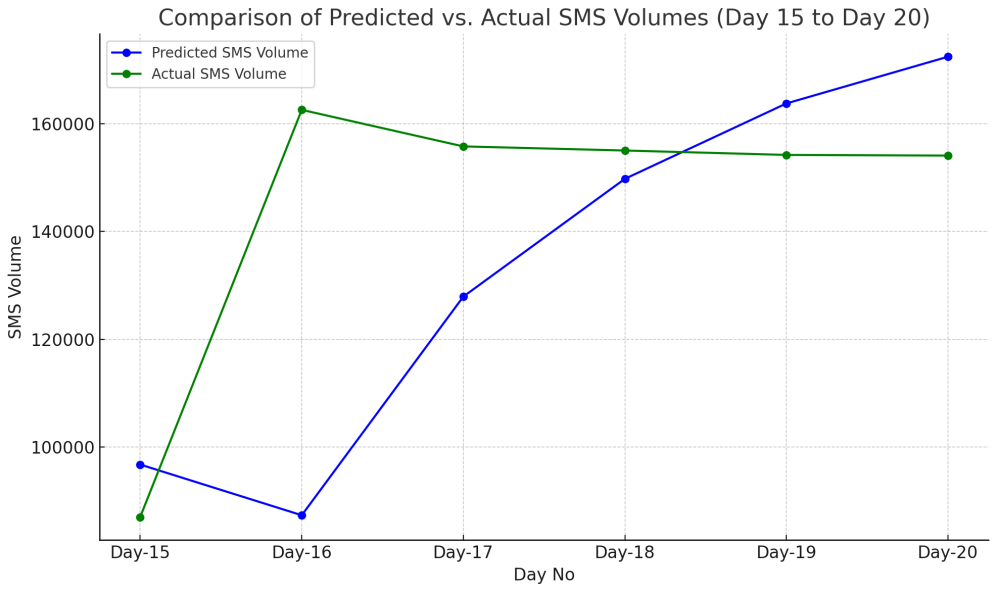

Figure 3.1: Visualization of Predicted vs. Actual SMS Volumes

The chart below visualizes the comparison between predicted and actual SMS volumes for Days 15 to 20, along with the error percentage for each day: Here is the chart that compares the predicted and actual SMS volumes for Days 15 to 20. The blue line represents the

predicted SMS volumes, while the green line represents the actual SMS volumes. This visualization clearly highlights where the predictions were close to the actual values and where they diverged.

Table 3.1: Comparison of Predicted and Actual SMS Volumes with Accuracy Percentage

Day No	Predicted SMS Volume	Actual SMS Volume	Error Percentage	% of Accuracy
Day-15	96719	86970	0.112096	90%
Day-16	87303	162602	-0.463088	186%
Day-17	127912	155813	-0.179067	122%
Day-18	149789	155048	-0.033919	104%
Day-19	163796	154230	0.062024	94%
Day-20	172475	154108	0.119183	89%

3.1.2 R-squared Analysis

The performance of our polynomial regression model was evaluated using the R-squared (R^2) value, a measure of how well the model explains the variance in the actual data. For our model, we obtained:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$R^2 = -0.7228$$

Where:

- y_i represents the actual SMS volumes.
- $\hat{y}_i$ represents the predicted SMS volumes.
- $\bar{y}$ represents the mean of the actual SMS volumes.
- The numerator captures the residual sum of squares (RSS), which measures the total error between the actual and predicted values.

- The denominator captures the total sum of squares (TSS), which represents the total variance in the actual data.

An R^2 value below zero indicates that the model performs worse than simply predicting the average of the actual values for all data points. This suggests that our model underfits the data, failing to capture the relationship between the date and SMS volume.

Implications

- The model does not adequately represent the trends in the data.
- Adjustments such as increasing the polynomial degree or using regularization techniques may improve performance.

Conclusion

The current model requires refinement, as indicated by the negative R-squared value. Further adjustments are necessary to better fit the SMS traffic data. By our understanding if we have 180 days real dataset we could have more accuracy.

Chapter 4

Conclusion and Future Works

4.1 Conclusion

This report presents a comprehensive study on forecasting SMS traffic and balance availability using machine learning techniques. By leveraging historical data from Wintel Limited, we developed a predictive model capable of providing accurate forecasts for SMS traffic and balance levels over the next two days. This model has demonstrated its ability to anticipate potential service interruptions, particularly during holidays and other peak periods, ensuring clients can optimize their SMS usage and avoid disruptions.
Key benefits of this work include:

- Improved service reliability during peak traffic periods.
- Enhanced resource allocation based on accurate demand forecasts.
- Increased customer satisfaction through proactive management of SMS traffic and balances.

Overall, this project contributes to the practical application of machine learning in managing communication resources, especially in time-sensitive environments where uninterrupted service is critical.

4.2 Future Works

While the current model demonstrates strong predictive capabilities, several areas for future research and improvement have been identified:

- **Model Optimization:** Future iterations of the model can explore the use of advanced machine learning techniques, such as deep learning or ensemble models, to improve prediction accuracy, especially for more complex traffic patterns. Also we don't have currently all 180 days real dataset to actually verify original outcomes. After getting 180 days dataset we can verify the model more accurately.

- **Incorporation of Additional Features:** Additional factors, such as external events, marketing campaigns, or changes in client behavior, can be incorporated into the model to refine its forecasts further.

- **Real-Time Prediction:** While this report focused on hourly predictions, the model can be extended to provide real-time updates, allowing for even more dynamic resource management.

- **Scalability:** Further research is needed to test the scalability of the model across a broader set of clients and operators, particularly for large-scale applications where SMS traffic volume is significantly higher.

- **Exploring Different Algorithms:** Polynomial regression was used for this project, but future work could evaluate the performance of alternative algorithms, such as decision trees, random forests, or gradient boosting methods, to find the best fit for the problem domain.

- **Integration with Automated Systems:** Finally, future work can focus on integrating the predictive model into automated systems that can automatically allocate resources or send alerts to clients when balances are predicted to drop below a critical threshold.

By addressing these areas, we can further enhance the reliability and robustness of SMS traffic and balance prediction systems, ultimately improving client experiences and operational efficiency.