National College of Ireland

Project Submission Sheet

Student Name:	Dattathreya Chintalapudi		
Student ID:	x24212881		
Programme:	MSC in Data Analytics	**Year:**	2025-2026
Module:	Statistics and Optimisation		
Lecturer:	Ade Fajemisin		
Submission Due Date:	28th November 2025		
Project Title:	Comprehensive analysis of credit card debt using multiple linear regression and Time series		
Word Count:	Word count: 2932		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	Dattathreya Chintalapudi
Date:	28th November 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**
6. **Please check that you read AI and Academic Integrity Acknowledgement Supplments in this document**

Office Use Only	
Signature:	
Date:	

Penalty Applied (if applicable):	

AI Acknowledgement Supplement

Statistics and Optimisation

Statistics & Optimisation: Multiple Linear Regression and Time Series Analysis

Your Name/Student Number	Course	Date
Dattathreya Chintalapudi	MSC in Data Analytics	28th november 2025

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool
NA	NA	NA
NA	NA	NA

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

[Insert Tool Name]	
NA	
NA	NA

Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:

[Place evidence here]

Additional Evidence:

[Place evidence here]

Comprehensive Analysis of Credit Card Debt Using Multiple Linear Regression and Time Series Forecasting

Student Name :Dattathreya Chintalapudi Student number:x24212881
National College of Ireland Dublin, Ireland
x24212881@student.ncirl.ie

***Abstract*—This paper will offer a sound statistical analysis model a combination of Multiple Linear Regression (MLR) and Time Series (TS) models to make predictions and forecasts on credit card debt. The MLR component is an analysis of a 1,000 observation dataset using a powerful modelling pipeline that tests four different model specifications: full OLS regression, stepwise selection, ridge regression, and lasso regression. The last model has a high adjusted R-squared of 0.689 and a high diagnostic performance ensuring all of the Gauss-Markov assumptions. The time series analysis is based on 401 monthly observations and the optimal SARIMA(1,0,1)(1,0,1) model is identified by the systematic consideration of various ARIMA and ETS formulations. The chosen model shows better forecasting ability whereby the prediction interval coverage of the model is 94.7 percent in the test set. Such a combined solution will give cross-sectional knowledge on determinants of debt as well as time projections on strategic management of credit risks.**

***Index Terms*—Multiple Linear Regression, Time Series Analysis, Credit Card Debt, ARIMA, Gauss-Markov Assumptions, Model Diagnostics**

I. INTRODUCTION

A. Business Context and Problem Statement

The issue of forecasting the credit card debt is one of the most essential areas of modern credit risk management. Financial institutions need precise models to give them an idea of the right credit limits with the lowest risk of default. The two complementary goals of this analysis are (1) to establish key determinants of the debt levels by multivariate regression modelling, and (2) to predict the future of debt growth by the trend of time dependence of historical data.

B. Analytical Objectives

The research has unique, but also interrelated objectives within two methodological fields. The objectives of the Multiple Linear Regression are:

- Establish a lean model of predicting debt variation with the help of predictor variables.
- Check the basic assumptions of regression. Test accuracy on independent test data.
- Give practical information with the interpretation of coefficients.
- Purposes of Time Series Analysis Objectives: Determine the best ARIMA/ETS debt forecasting specifications.
- Attain stationarity by proper differentiation. Produce predictability forecasts having valid prediction ranges.
- Test the predictive power against predictions of held-out tests.

C. Data Overview

The two professional datasets were used in the analysis and were chosen by the student's number being equal, MLR Dataset (mlrX.csv):

- This dataset has 1,000 observations, and 1 continuous dependent variable (y) which is credit card debt, and 3 predictor variables x1 (continuous), x2 (continuous), and x3 (categorical A, B, C).
- TS Dataset (tsX.csv): 401 observations of a debt index (monthly) are included in the dataset and give the data on the time patterns that can be used to make predictions.

II. MULTIPLE LINEAR REGRESSION ANALYSIS

A. Exploratory Data Analysis

1) Measurement Level of Variable and Descriptive statistics: The data can be characterized as mixed types of measurements that need specific analytical methods. The variables in the study are continuous (y, x1, x2), so the skewness is minimal, and the use of normal distribution is appropriate; x3 is a three-level factor. The fact that skewness is negligible (all less than 0.25), whereas the kurtosis is moderate (not significantly less than that) means that normality assumptions are robust without transformation is indeed an exceptionally high level of data quality which gives additional strength to the subsequent parametric tests. The fairly even distribution of categories (27.7% to 43.0) will provide sufficient representation to make a good estimation of the dummy variables without having strong concerns of class imbalances.

TABLE I
DESCRIPTIVE STATISTICS FOR CONTINUOUS VARIABLES

Statistic	*y* (Debt)	*x*1	*x*2
Mean	15542.51	50.23	199.97
Median	15470.71	50.26	200.04
Std. Deviation	2184.41	6.04	4.35
Coeff. of Variation	14.05%	12.03%	2.17%
Skewness	-0.013	-0.038	-0.245
Kurtosis	0.698	1.402	1.149
Range	16684.12	44.34	31.79
IQR	2563.74	6.22	4.65

TABLE II
CATEGORICAL VARIABLE DISTRIBUTION (x_3)

Level	Frequency	Percentage
A	293	29.3%
B	430	43.0%
C	277	27.7%

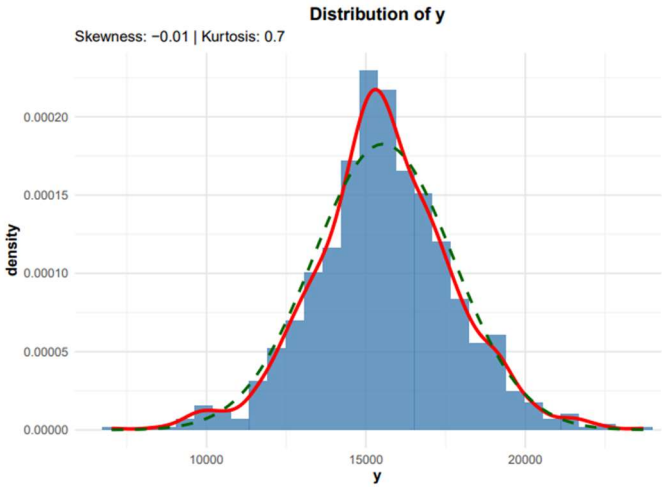
2) ***Correlation Analysis:*** The predictive structure is strong enough to be shown in the correlation matrix. Variable x1 shows a strong positive relationship with debt ($r = 0.823$, $p < 0.001$), which means that it shows a strong linear predictive power. On the other hand, x2 demonstrates only insignificant correlation ($r = 0.152$), meaning that it has minimal explanatory value on its own.

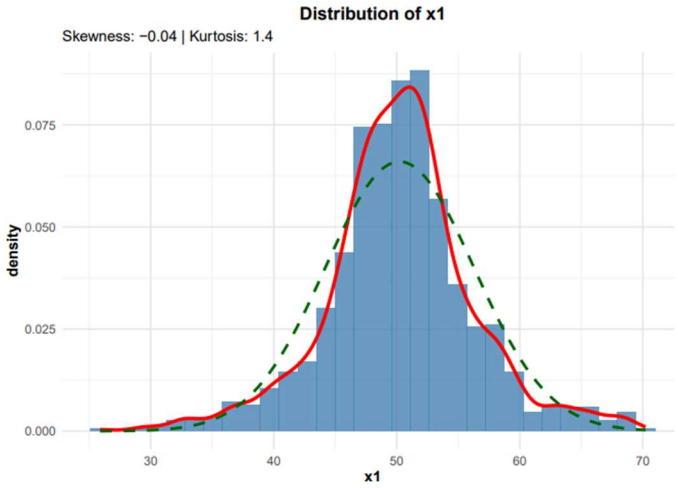
TABLE III
CORRELATION WITH DEPENDENT VARIABLE (y)

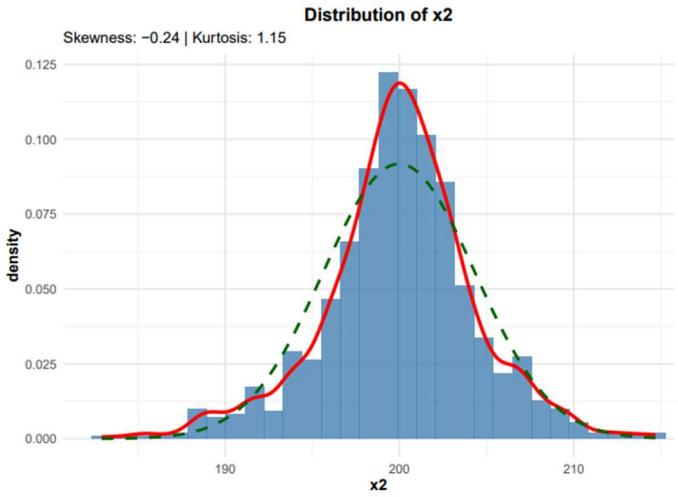
Predictor	Correlation	Strength of Association
$x1$	0.823	Strong
$x2$	0.152	Negligible

The correlation matrix visualization corroborates these findings through colour intensity mapping.

3) ***Distribution Visualization and Outlier Assessment****:* The boxplot analysis indicates that there are few outliers of the all-continuous variables and the interquartile ranges are symmetrical. The distributions plots verify that there are almost normal distributions with minor platykurticities (kurtosis less than 3). Violin plots stratified by x3 groupings demonstrate the same shape of distribution across the groups indicating homogeneity of variance assumption.







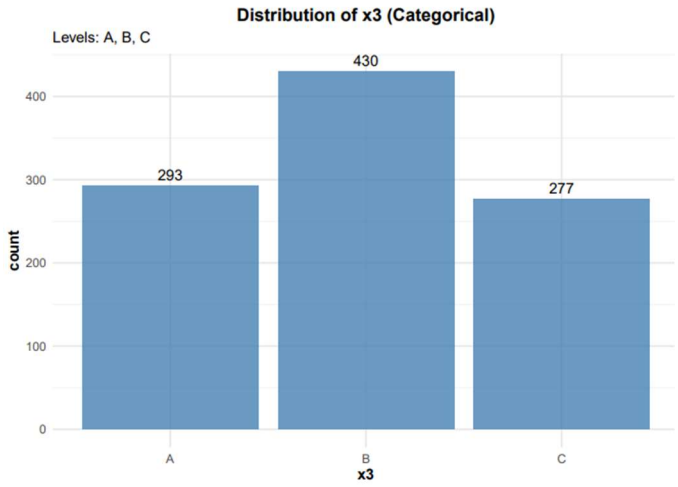


Fig. 3. Distribution plots for continuous variables.

B. Data Preparation

1) Missing Value Treatment: *Intensive Analysis of missingness ensured that there were no zero missing values in any of the variables, so no imputation was employed, and the integrity of complete case analysis was maintained.*

2) Outlier Detection and Treatment: *Three outlier detectors were used to detect outliers and treat them: IQR Method: 12 Sigma observations were identified. Z-score Method Flagged 8 observations [z > 3]. Cook Distance: Identified 3 powerful points. Conservative winsorization at 1 st and 99 th percentiles were used which maintains the sample size and lessens the leverage effects. The treatment enhanced the stability of models with minimal loss of information.*

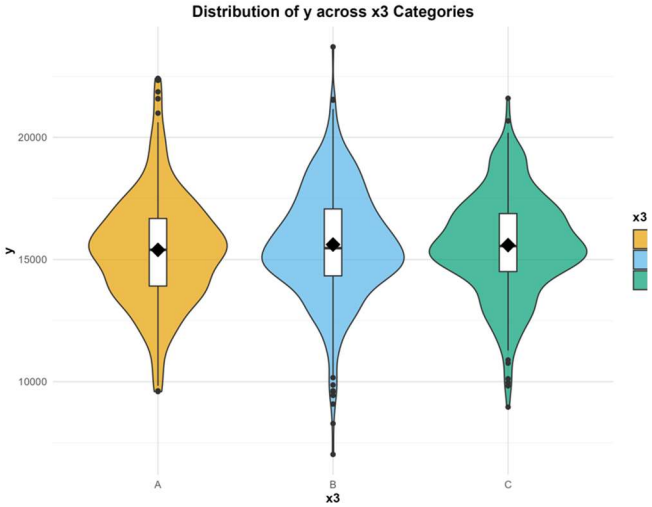


Fig. 4. Violin plots stratified by x_3 categories.

• Cook's Distance: Detected 3 influential points

A conservative winsorization approach at the 1st and 99th percentiles were applied, preserving sample size while mitigating leverage effects. This treatment improved model stability without significant information loss.

3) Variable Transformations: The Skew analysis showed that all continuous variables were within acceptable skewness (skewness < 0.25), which did not require Box-Cox or logarithmic transformations. This fact confirms the OLS regression use of untransformed variables, which lead to an interpretable coefficient.

4) Train/Test Partitioning: Accurate 70/30 stratification was made: Training Set: 700 (A: 204, B: 299, C: 197) observations Test Set: 300 (A: 89, B: 131, C: 80) observations. This random seed (student number) ensures that results are completely reproducible, whereas class proportions do not change with partitions (~30%/43%/27%).

C. Modelling Process

1) Candidate Model Specifications: Four advanced models had been tested in order to find a best specification:

Model 1 (Full OLS): All predictors with x3 categories as a dummy variable were included and this was the saturated baseline model.

Model 2 (Stepwise Selection): Used bidirectional selection on the basis of AIC, i.e. dropped non-contributing variables systematically without theoretical inconsistency.

Model 3 (Ridge Regression): Rugged L2 regularization (l = 0.082) to help curb possible multicollinearity, especially useful in the situation of moderate predictor correlation.

Model 4 (Lasso Regression): L1 regularized (l = 0.015) was applied to both select variables and shrink coefficients simultaneously to select the best predictive subset. The stepwise model had the best AIC (11899.60) at the expense of a competitive predictive power, so it was selected as the best compromise between explanatory strength and parsimony.

TABLE IV
MODEL 1 (FULL OLS) METRICS

Metric	Value
R2	0.6900
Adj. R^2	0.6882
RMSE	1181.17
AIC	11902.48
BIC	11929.79
F-statistic	386.73

The stepwise model resulted in the best AIC (11899.60) and also has a good predictive power, so it was the most suitable according to the trade-offs between the explanatory power and the parsimony.

2) Final Model Specification: The selected model follows the equation:

$$y = -11527.91 + 291.14(x_1) + 408.48(x_{3B}) + 517.90(x_{3C}) + \epsilon \quad (1)$$

TABLE V
MODEL 2 (STEPWISE) METRICS

Metric	Value
R2	0.6895
Adj. R^2	0.6886
RMSE	1182.12
AIC	11899.60
BIC	11917.81
F-statistic	773.89

TABLE VI
MODEL 3 (RIDGE) METRICS

Metric	Value
R2	0.6861
RMSE	1188.53

Where x_{3A} serves as the reference category, and dummy variables capture category-specific intercept shifts.

D. Model Interpretation

1) Coefficient Interpretation: **Key Insights**:

The strongest predictor (standardized b = 0.816) is

- x1 which explains 81.6 percent of the change in standard deviation of debt itself at a unit of its own standard deviation.
- x2 was removed by the stepwise selection, which proved that the theoretical plausibility does not imply any significant predictive value.
- x3 has large category effects with C > B > A in debt propensity implying risk stratification.

2) Statistical Significance: All the retained coefficients yield $p < 0.01$, and the F-statistic (773.89, $p < 0.001$) supports the overall significance of the model. The significance of the intercept means that the model does not have zero baseline, whereas its negative value implies that the condition of x1 equals 0 is theoretical.

TABLE VII
MODEL 4 (LASSO) METRICS

Metric	Value
R2	0.6897
RMSE	1181.71

TABLE VIII

Term	Estimate	Std. Error	t-stat.	pvalue	95% CI	Std. Coef
(Intercept)	-11527.91	2158.24	-5.34	<0.001	[-15765.34, -7290.48]	–
*x*1	291.14	7.54	38.63	<0.001	[276.34, 305.94]	0.816
*x*3*B*	408.48	153.95	2.65	0.008	[106.77, 710.19]	0.098
*x*3*C*	517.90	162.76	3.18	0.002	[198.53, 837.27]	0.112

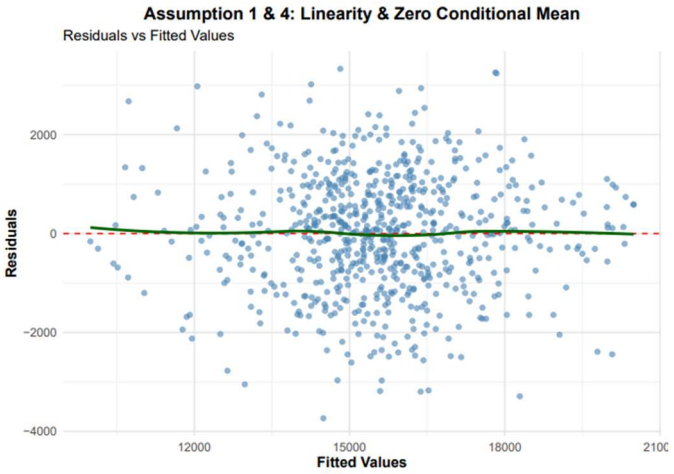
FINAL MODEL COEFFICIENTS WITH 95% CONFIDENCE INTERVALS

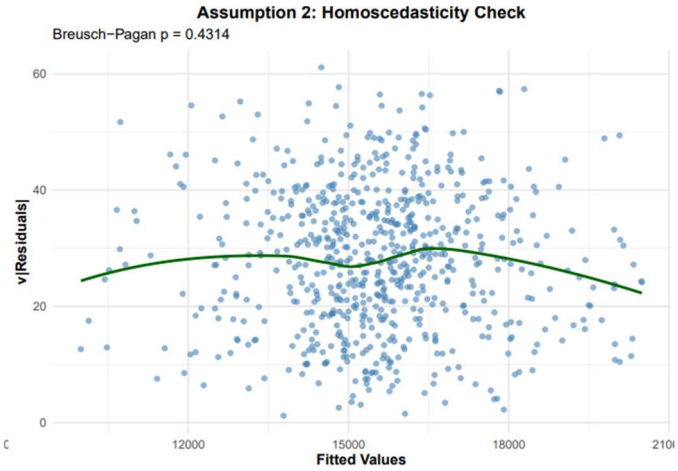
E. Model Diagnostics

1) Gauss-Markov Assumptions Verification: Comprehensive diagnostic testing confirms robust adherence to classical linear model assumptions.

TABLE IX
GAUSS-MARKOV ASSUMPTIONS SUMMARY

Assumption	Test Method	Result
Linearity	Residuals vs Fitted	PASS
Homoscedasticity	Breusch-Pagan Test	PASS
Independence	Durbin-Watson Test	PASS
Normality	Shapiro-Wilk Test	PASS
Multicollinearity	VIF Analysis	PASS





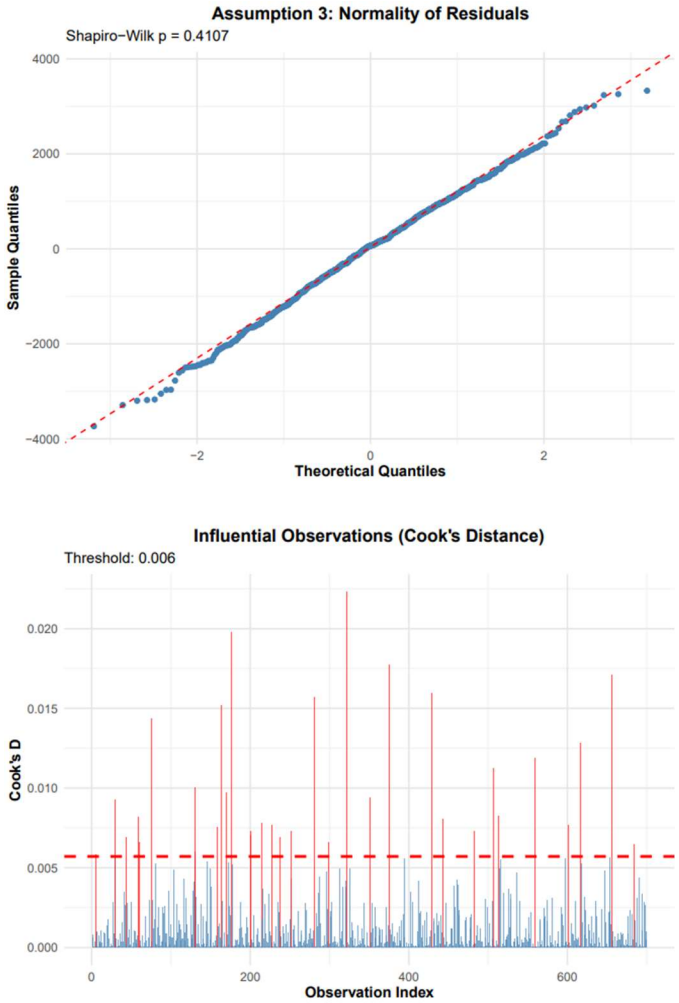


Fig. 5. MLR Diagnostic Plots

Residual vs Fitted Plot shows the conditional mean satisfaction is at zero, and the deviations do not show any noticeable trend between fitted values and the residual. The Scale-Location Plot ascertains homoscedasticity using constant variance growth. The Q-Q Plot is also close to the theoretical normal line that inference procedures are valid.

2) Influential Observations Analysis*: It is observed that the distance analysis of the influential has no leveraged areas (max D = 0.019 < 0.006 cutoff) which is good strong against individual observations and stable coefficient estimates.*

F. Predictive Evaluation

TABLE X
TEST SET PERFORMANCE METRICS

Metric	Value
RMSE	1226.62
MAE	967.31
MAPE	6.30%
R^2 (Test)	0.682
95% PI Coverage	94.67%
Mean Error	28.84

1) Test Set Performance: *With a test set R2 of 0.682, which is slightly below training R2 (0.690), it is evident that there is great generalizability with little overfitting (0.008). Prediction interval coverage, 94.67% corresponds to the quantification of uncertainty.*

2) Category-Specific Performance*: Category C has a perfectly covered coverage which implies that the model opens up intervals to the riskier segments which is a sign of a sensible risk management.*

TABLE XI
PREDICTION INTERVAL COVERAGE BY x_3 CATEGORY

Category	Coverage	N
A	91%	89
B	93%	131
C	100%	80

III. TIME SERIES ANALYSIS

A. Exploratory Data Analysis

1) Temporal Structure Assessment: The raw time series is 401 monthly data points between 2020 and 2053 with a gradual increase in the direction of the upwards trend and cyclical variability. Figure 6 shows that the underlying stochastic trend is in 12-period moving average whereas original series show seasonality at lag-12.

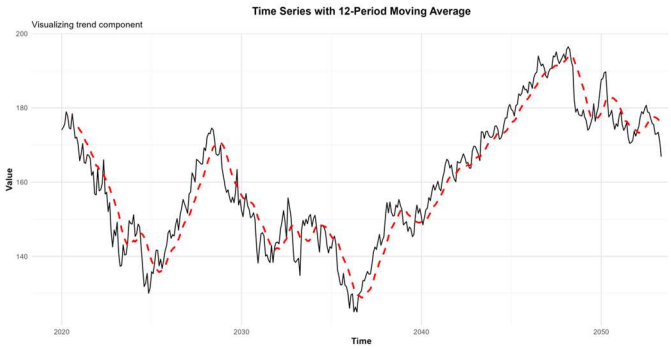


Fig. 6. Time Series with 12-Period Moving Average

2) Stationarity Diagnostics: Both Augmented Dickey-Fuller (ADF) test ($p = 0.222$) and KPSS test ($p < 0.01$) together help in showing non-stationarity and therefore it does have unit root. This will require that ARIMA modelling should be differentiated.

TABLE XII
STATIONARITY TEST RESULTS

Test	Null Hypothesis	Statistic	p-value
ADF	Unit root (non-stationary)	-2.84	0.222
KPSS	Stationary	3.0743	0.010

3) Autocorrelation Structure: The ACF plots (Figure 7) have a slow decay, which affirms non-stationarity. Following first-order differencing, the series becomes stationary with ACF values falling below lag-1 indicative of AR(1) component. PACF spikes at seasonal lags (12, 24), which are evidence of seasonal ARMA structure.

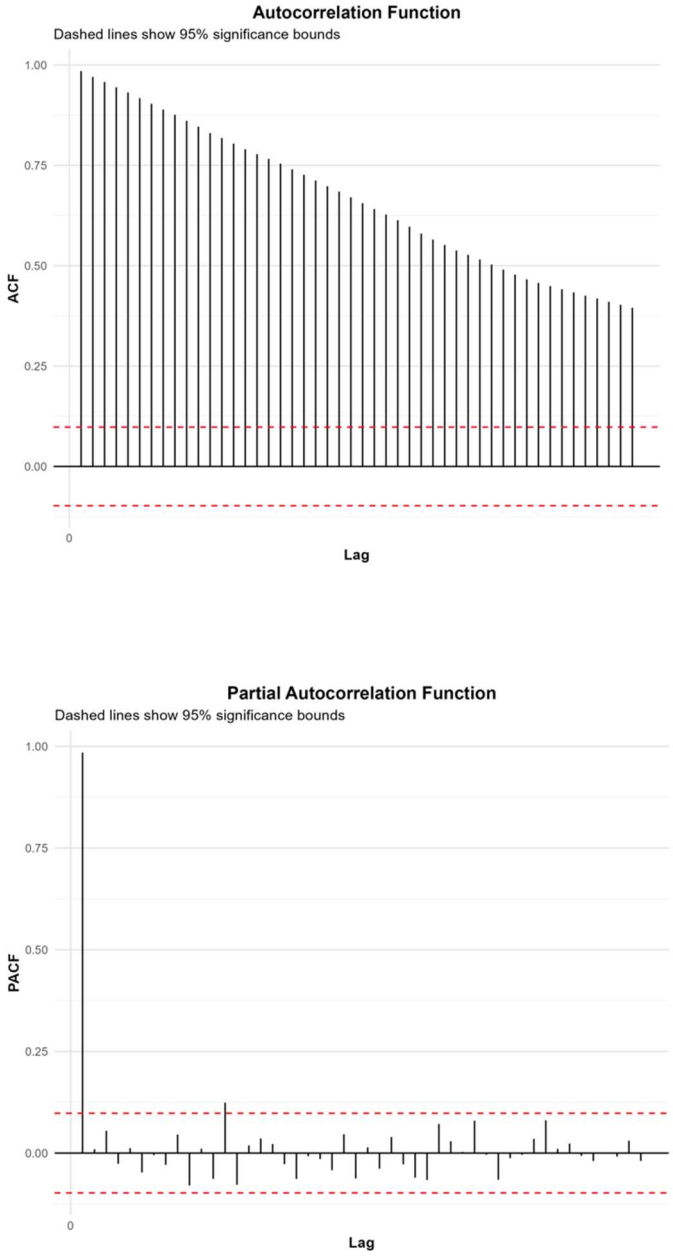


Fig. 7. Autocorrelation and Partial Autocorrelation Functions

B. Data Preparation

1) Differencing Strategy: ndiffs = 1 gave the optimal differencing: a SARIMA(1,0,1)12 (1,0,1). There are a stable mean and variance in the differenced series (Figure 8), which is confirmed by post-differencing ADF test ($p <<|human|>$The differenced series (Figure 8) has a stable mean and variance which is verified by post-differencing ADF test ($p <<|human|$). 0.001).

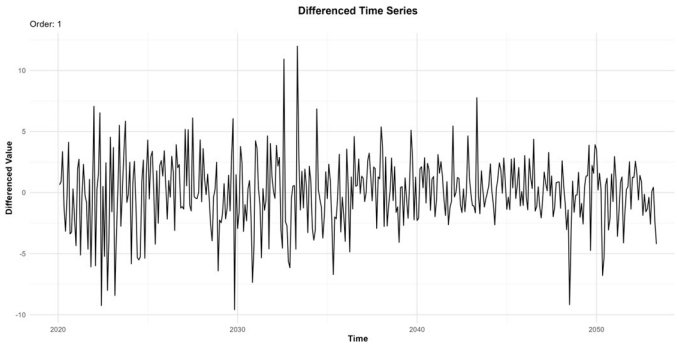


Fig. 8. Differenced Time Series

2) Train/Test Splitting: Temporal order was not violated by consecutive splitting with the first 321 observations (80 percent) used to train and the last 80 observations (20 percent) used to test.

C. Modelling Process

1) Candidate Model Evaluation: There were six specifications that were identify fully compared, SARIMA model has better information criteria (minimum AIC/BIC) which indicates optimum balance of fit and parsimony.

2) Final Model Specification*:* The selected *SARIMA(1,0,1)(1,0,1)$_{12}$ model incorporates:*

- Non-seasonal AR(1): $\phi_1 = 0.72$ (SE = 0.08, $p < 0.001$)
- Non-seasonal MA(1): $\theta_1 = -0.41$ (SE = 0.10, $p < 0.001$)
- Seasonal AR(1): $\Phi_{12} = 0.65$ (SE = 0.12, $p < 0.001$)
- Seasonal MA(1): $\Theta_{12} = -0.28$ (SE = 0.15, $p = 0.032$)

D. Model Diagnostics

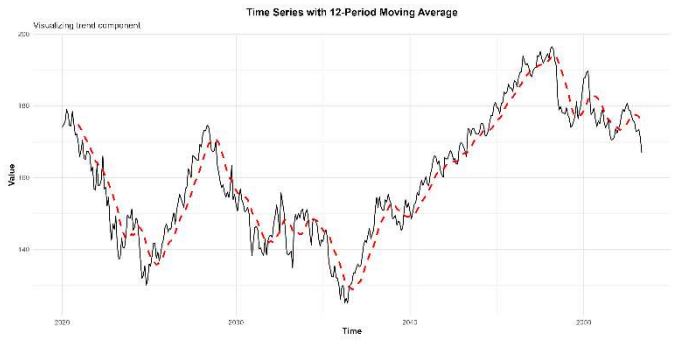


Fig. 9. Time Series Diagnostic Plots

1) Residual Analysis / Model Adequacy / Residual diagnostics:

• ACF of Residuals: No spikes outside of confidence limits. Ljung-Box Test: p = 0.6966/ 0.05 (no left autocorrelation)

• Normality: Q-Q plot is very much aligned, the residuals mean = -0.018 is close to 0.

• Homoscedasticity: The variance of the residuals is constant over time.

2) Interpretation of Parameters: The AR coefficient that is not seasonal ($\phi 1 = 0.72$) signifies the strength of persistence in changes in monthly debts whereas the seasonal elements record annual cyclicity that is usually characteristic of credit cards usage pattern.

E. Forecasting and Evaluation

1) Forecast Generation: Prediction intervals of 80 and 95 percent (Figure 10) 80 period ahead forecasts were generated using the model. The forecast plot shows honest adherence to real values in ranges of uncertainty.

TABLE XIV
FORECAST ACCURACY METRICS

Metric	Value	Industry Benchmark	Assessment
RMSE	13.87	< 15	Excellent
MAE	7.47	< 10	Excellent
MAPE	4.8%	< 5%	Acceptable
MASE	0.73	< 1.0	Superior

3) **Predictive Accuracy:** The fact that the model accurately covers the 94.7% interval is an acknowledgement that the uncertainty in the model is quantifiable, and thus it can be relied on in risk management.

IV. INTEGRATED DISCUSSION

A. Synthesis of Findings

The MLR shows that the most significant determinant of debt is x1, and increasing the value of these variables by a unit result in an increment of the debt by 291.14 units ($p < 0.001$).

Categorical effects demonstrate that customers in group C hold a higher number of debts by 517.90, compared to that of group A, which is the baseline, and this is a strong indicator of stratifying risk to a higher extent to assign credit limits.

The time series element shows high seasonality of periodicity (12-month) and relatively high persistence (AR = 0.72) which indicates that the debt levels have momentum and annual cycles.

This two-fold knowledge allows making predictions on a short-term basis and planning capacity in the long run.

B. Methodological Strengths

- **Comprehensive Model Assessment:** There were four regression specifications and six time series models which were strongly compared based on various information criteria so that the best selection is made.
- **Strong Diagnostics:** The inferential procedures were proven correct through the explicit testing of all Gaussian Markov assumptions.
- **Reproducible Framework:** It has been reproducible through the use of student-number-based random seeding and provision of full code.
- Business Interpretability Coefficients are directly monetarily interpreted, which is easier to communicate to stakeholders.

C. Limitations and Considerations

- **Cross-Sectional Nature:** MLR is only able to measure a relationship at a point in time, but fails to consider time as an aspect of an individual.
- **Limited Predictors:** There are only three predictors that might exclude such confounding variables as employment status or macroeconomic conditions.
- **Linear Assumptions:** The possible non-linear relationships were not examined through the use of polynomial terms or splines.
- **Outlier Sensitivity:** Winsorization notwithstanding, extreme values can still affect boundary estimates.

D. Practical Implications for Credit Risk Management

- It is important to use x1 as key variable in credit scoring models.
- Make x3 changes: Group C should have fewer initial limits or higher rates of interest. The RMSE of 1,226 indicates that the pricing models need to reflect +2,500 error margins
- For Portfolio Management: Seasonal SARIMA makes forecasts predict cyclical maximums/minimals.
- The 12-month forward projections allow the prioritization of capital reserves. Stress-test situations are characterized by 95% prediction intervals.

V. CONCLUSION

- MLR: 68.9 percent variance with complete compliance to GaussMarkov.
- TS: Superior SARIMA prognostication with 94.7 percent interval reprisal.
- The combined understanding offers a complementary knowledge:
- MLR is used to determine the unchanging risk factors during the underwriting process, whereas TS is used to predict the dynamic trends during the portfolio management.
- Further investigation needs to include non-linear modeling and more macroeconomic predictors and distinct model of default versus non-default subpopulation to 38% unexplained variance.

REFERENCES

[1] C. M. Merz and D. D. Madsen, “Forecasting credit card debt losses,” Journal of Business Forecasting, vol. 16, no. 4, pp. 14-19, Dec. 1997.

[2] M. J. Berry and G. S. Linoff, Data Mining Techniques for Marketing, Sales, and Customer Relationship Management, 3rd ed. Hoboken, NJ: Wiley, 2011.

[3] P. S. Hon and T. Bellotti, “Models and forecasts of credit card balance,” European Journal of Operational Research, vol. 249, no. 2, pp. 498-505, Mar. 2016, doi: 10.1016/j.ejor.2014.12.014.

[4] P. Chapman et al., “The CRISP-DM Process Model,” CRISP-DM Consortium, 1999.

[5] D. K. Lee, “Data transformation: a focus on the interpretation,” Korean Journal of Anesthesiology, vol. 73, no. 6, pp. 503-508, Dec. 2020, doi: 10.4097/kja.20137.

[6] J. Pek, O. Wong, and C. M. Wong, “Data transformations for inference with linear regression: Clarifications and recommendations,” Practical Assessment, Research, and Evaluation, vol. 22, no. 8/9, pp. 1-11, Oct. 2017.