ELLER COLLEGE OF MANAGEMENT
HSLopez School of Business Analytics

ANALYTICS REPORT

TO: SOFI MARKETING TEAM

FROM: JAKE MOORE

SUBJECT: ANALYSIS OF CLIENT DEBT

DATE: OCTOBER 16TH, 2025

Introduction

In this brief analysis, we analyzed data from approximately 2,000 SoFi clients to determine how Zip Code Per Capita Income, Client Yearly Income, Current Age, and Number of Credit Cards best predict Total Debt. The purpose of this analysis is to determine which of these variables significantly impacts total debt so that SoFi can better target prospective clients for debt consolidation services.

A multiple regression analysis was conducted using four variables: Zip Code Per Capita Income, Client Yearly Income, Current Age, and Number of Credit Cards. The analysis included tests for significance and evaluations of the model fit using R^2 and standard error. To ensure the reliability of our results, we examined residual plots for potential violations of assumptions, such as heteroskedasticity, and tested for multicollinearity and endogeneity.

The results show that all four predictors are statistically significant. Clients who are younger, have fewer credit cards, higher personal income, and live in lower-income zip codes tend to have higher total debt. Although the model had its limitations, as it has a standard error of $41,300, and showed possible multicollinearity and heteroskedasticity for our income-related variables.

Based on our findings, we recommend using the model as a tool to identify clients with higher amounts of debt, while acknowledging the individual-level prediction limitations. SoFi should be cautious, as the model has a high amount of variability in accuracy.

Data Analysis

The following data analysis section presents regression equations, an evaluation of the model fit, and multiple hypothesis tests used to understand how Total Debt is influenced by Zip Code Per Capita Income, Client Yearly Income, Current Age, and Number of Credit Cards. Additional tests were also used to evaluate assumptions, including residual plots, multicollinearity, and endogeneity. Finally, we include an example scenario of how the model can be used to predict the total debt of a hypothetical client, followed by a recommendation for how SoFi could use the findings to guide marketing campaigns.

Population Regression Equation

$$Total\ Debt = \beta_0 + \beta_1(Zip\ Code\ Income) + \beta_2(Client\ Yr\ Income) + \beta_3(Age) + \beta_4(Number\ of\ CC) + \varepsilon$$

Estimated Sample Regression Equation

$$\widehat{Total\ Debt} = 44.69 - 0.66(Zip\ Code\ Income) + 1.51(Client\ Yr\ Income) - 0.55(Age) - 3.20(Number\ of\ CC)$$

Fit of the Model

R^2 Interpretation: 0.377 tells us we are 37.7% of the way toward perfectly predicting total debt using this model.

Standard Error Interpretation: The standard error is 41.3, which is represented in thousands of dollars. This standard error value represents that, on average, the model's prediction of total debt differs by about $41,300.

Significance of Variables

The following variables are significant predictors of Total Debt because they have p-values less than 0.05: Zip Code Per Capita Income, Client Yearly Income, Current Age, and Number of Credit Cards.

Coefficient Interpretations

Zip Code Per Capita Income:
As zip code per capita income increases by $1,000, total debt decreases by $644, on average, and all else constant.

Client Yearly Income:
For every additional $1,000 of the client's yearly income, total debt decreases by $1,509, on average, and all else constant.

Current Age:
For every year increase in the client's current age, total debt decreases by $546, on average, and all else constant.

Number of Credit Cards:
As the number of credit cards increases by one, total debt decreases by $3,202, on average, and all else constant.

Residual Plot Interpretations

Zip Code Per Capita Income:
This residual plot appears to have a funnel shape. This indicates possible issues with heteroskedasticity/changing variability. This could cause the standard error for Zip Code Per Capita Income to not be correct meaning the p-value for Zip Code Per Capita Income could also be incorrect. We would fix this by using White's Standard Errors.

Client Yearly Income:
This residual plot appears to have a funnel shape. This indicates possible issues with heteroskedasticity/changing variability. This could cause the standard error for Client Yearly Income to not be correct meaning the p-value for Client Yearly Income could also be incorrect. We would fix this by using White's Standard Errors.

Current Age:
This residual plot appears to have a random shape. This indicates no issues with non-linear patterns or heteroskedasticity/changing variability. So far, we can trust the p-value associated with Current Age.

Number of Credit Cards:
This residual plot appears to have a random shape. This indicates no issues with non-linear patterns or heteroskedasticity/changing variability. So far, we can trust the p-value associated with Num Credit Cards.

Check for Multicollinearity and Endogeneity

Multicollinearity:
We do have evidence of multicollinearity between Zip Code Per Capita Income and Client Yearly Income because the correlation of 0.96 is larger than 0.8. This could have caused one or both of these variables to not be reported as significant even if they really were, but they were both still reported as significant in this model. While we could try to fix this issue, we don't necessarily need to since both variables are already significant.

Endogeneity:
All the correlations between the x-variables and the residuals are extremely close to zero. This indicates we do not have issues with endogeneity. This means we can trust the coefficients estimated in the regression.

Example Prediction
Using our regression model, we can predict the total debt for a 28-year-old client who earns $45,000 per year, lives in a zip code with a per capita income of $40,000 and has two credit cards.

Intercept = 44.69
Zip Code Income = -0.66 x 40 = -26.40
Client Income = 1.51 x 45 = 67.95
Age = -0.55 x 28 = -15.40
Number of Credit Cards = -3.20 x 2 = -6.40

$$\widehat{Total\ Debt} = 44.69 - 0.66(40) + 1.51(45) - 0.55(28) - 3.20(2)$$
$$44.69 - 26.40 + 67.95 - 15.40 - 6.40 = 64.44$$

The predicted total debt is approximately $64,440. While this model is a methodological approach to estimating debt, we have to note that the standard error of $41,300 could cause this prediction to vary quite drastically for individual clients.

Recommendation

Based on the regression results, Client Yearly Income, Zip Code Per Capita Income, Current Age, and Number of Credit Cards are statistically significant predictors of Total Debt. Due to these relationships, we recommend that the SoFi debt consolidation team use this model to identify clients with the potential of having higher debt, specifically those with higher personal income, who are younger, have fewer credit cards, and live in lower-income zip codes. SoFi can leverage these different patterns to identify which clients might benefit from debt consolidation based on their demographic and financial profiles.

Conclusion

In this analysis, we examined data from approximately 2,000 SoFi clients to evaluate which factors best predict Total Debt. Our regression results showed that Zip Code Per Capita Income, Client Yearly Income, Current Age, and Number of Credit Cards are all statistically significant predictors. Specifically, clients with higher income, who are younger, have fewer credit cards, and live in lower-income zip codes tend to have higher total debt.

While the model highlights meaningful trends, it is important to note the potential limitations of the model, such as heteroskedasticity in income-related variables and a standard error of $41,300. These factors can greatly reduce the reliability and accuracy of the model at the individual level. Despite these limitations, the model can still be useful as a screening tool to find potential targets for their marketing campaigns.

Please feel free to contact me at jakemoore@arizona.edu if you have any questions or would like to discuss these recommendations in more detail.

Technical Appendix

Figure 1 – Regression Output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.613829746
R Square	0.376786957
Adjusted R Square	0.375537407
Standard Error	41.2930083
Observations	2000

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	4	2056625.77	514156.4424	301.5381284	5.3224E-203
Residual	1995	3401699.507	1705.112535		
Total	1999	5458325.277			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	44.68503295	3.354312561	13.3216664	7.51102E-39	38.10671011	51.26335578	38.10671011	51.26335578
Zip Code Per Capita Income ($1000s)	-0.663510009	0.331892467	-1.999171642	0.045725212	-1.314402183	-0.012617835	-1.314402183	-0.012617835
Client Yearly Income ($1000s)	1.508989968	0.164531105	9.171457091	1.12849E-19	1.186319166	1.831660769	1.186319166	1.831660769
Current Age	-0.545568329	0.061317708	-8.897402578	1.25274E-18	-0.665821784	-0.425314874	-0.665821784	-0.425314874
Num Credit Cards	-3.202434297	0.6448838	-4.965909048	7.41962E-07	-4.467150613	-1.937717981	-4.467150613	-1.937717981

Jake Moore

Figure 2 – Zip Code Per Capita Income Residual Plot

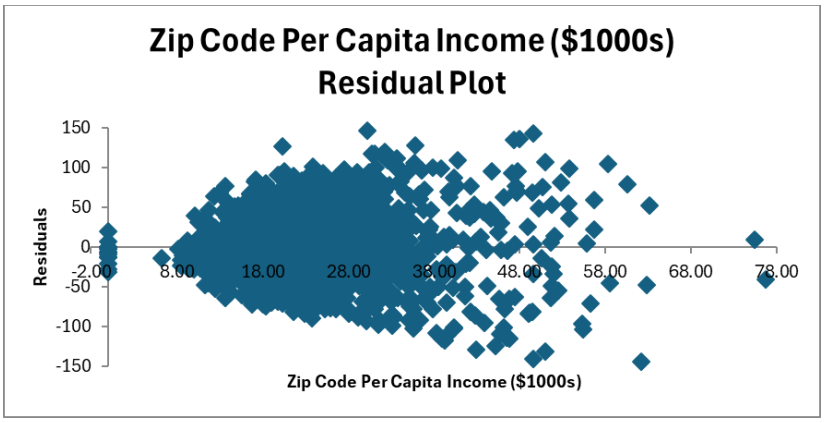

Figure 3 – Client Yearly Income Residual Plot

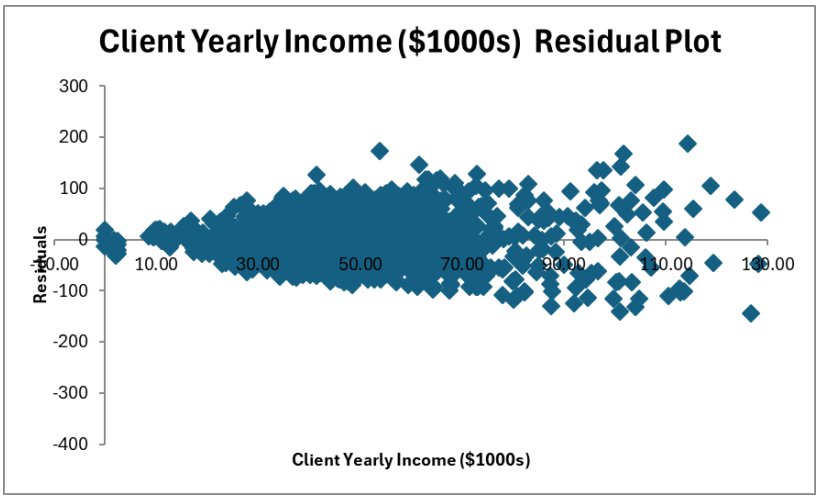

Figure 4 – Current Age Residual Plot

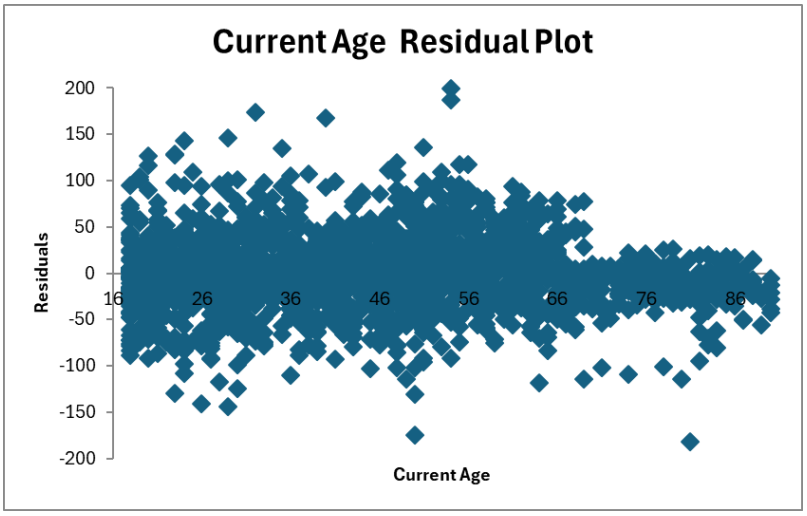

Figure 5 – Number of Credit Cards Residual Plot

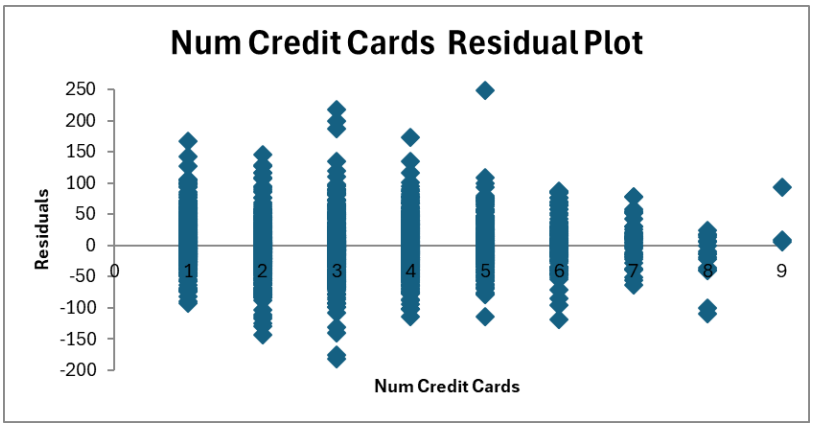

Figure 6 – Correlation Matrix

	Residuals	*Zip Code Per Capita Income ($1000s)*	*Client Yearly Income ($1000s)*	*Current Age*	*Num Credit Cards*
Residuals	1				
Zip Code Per Capita Income ($1000s)	1.12626E-15	1			
Client Yearly Income ($1000s)	1.08097E-15	0.963974613	1		
Current Age	8.86456E-17	-0.009053888	-0.114316421	1	
Num Credit Cards	2.31652E-16	0.018606403	-0.032876519	0.484189245	1
Jake Moore					