**Group Assignment 2**

**Total points -100**

**Clearly state your group number# in your submission files, including .R file. Name your .R file with your group number.**

**Interaction terms and indicator variables**

The following questions refer to the dataset “**Catalogs.csv**” posted on eCampus; download the data and save it to your hard drive. The data contains the information about the direct marketer discussed In class; in particular, the data contains information about the age (coded as young, mid or old) and gender of a customer, whether s/he owns a home or rents, whether or not s/he is married, whether or not s/he lives far from a brick-and-mortar store selling similar items, the customer’s salary, number of children, and purchase history (coded as low, medium, high and NA, if there is no history) and the number of catalogs we have sent that customer in the past. And finally, the data records the amount of money the customer has spent. The following questions refer to this data.

1. let’s investigate the joint impact of salary and the number of children on the amount spent; to that end, run a regression model with “AmountSpent” as the response variable and “Salary” and “Children” as the only 2 predictors. Submit R regression output including the lm formula equation screenshot, answer the following questions:
   1. Is “salary” an important predictor for amount spent? Why? (5 points)
   2. Do salary and number of children both have the same relationship with amount spent? (5 points)
2. First transform two variables “Amount Spend” and “Salary” into **log** forms. Create a scatterplot with regression lines for “Log.AmountSpent” as the response variable and “log.Salary” as the predictor, group the data points on scatterplot by “Location”. The scatterplot should have two colors, one for Location being “Far” and the other for location being “Close”. Draw regression line for each Far and Close. Submit your visual of the scatterplot
   1. State your observation on the nature of intercept and slope of these two regression lines. (5 points)
3. Next, let’s investigate the effect of purchase history on the amount a customer spends.
   1. Now, run a regression model with AmountSpent as the response variable and History as the (only) predictor. Provide a precise interpretation of the resulting coefficient labeled as “History [T.Low]” (5 points)
   2. Now, add the interaction term between salary and history to the above model; in other words, run a regression model with salary, history and the interaction term salary\*history (using amount spent as the response variable). Submit R regression output including the lm formula equation screenshot, answer the following questions:
      1. Is this model (including the interaction term) better than the previous model (from question 3.a)? (5 points)
      2. Provide a concise interpretation for the coefficient of Salary (5points)
      3. Provide a concise interpretation for the coefficient of the *interaction* between Salary and History[T.Low] ( 5 points)

**Handling Multicollinearity**

The following questions refer to the dataset “GrpHW2.StockPrice.csv” posted on eCampus; download the data and save it to your hard drive. The data contains financial indicators information for over seven thousand companies. Our goal is to develop a model to predict company stock price. To that end, we are going to focus only on those variables listed here:

|  |
| --- |
| 1. Stock Price 2. Total Debt 3. Cash 4. Revenues: Last yr 5. Invested Capital 6. Reinvestment Rate 7. ROE 8. ROC 9. Net Margin 10. Invested Capital 11. BV of Assets 12. Net Income 13. EBIT 14. EBITDA 15. FCFF 16. Cash as % of Firm Value 17. Cash as % of Revenues 18. Cash as % of Total Assets 19. Capital Expenditures 20. Depreciation 21. Trailing Revenues 22. Trailing Net Income 23. Intangible Assets/Total Assets 24. Fixed Assets/Total Assets 25. Market D/E 26. Market Debt to Capital 27. Book Debt to Capital |

1. First, investigate the data for correlation measures. State **4 variables** with correlation values above 0.6. State those variables correlation values and which variables they are correlated with. (5 Points)
2. Now, we would like to investigate the impact of all the variables on Stock Price: To that end, run a regression model with “StockPrice” as the response variable and all other variables as predictors; Submit R regression output including the lm formula equation screenshot, answer the following questions:
   1. Is this a good model? Why or why not? (5 points)
   2. Make **one suggestion** to improve this model. (5 points) \*Please keep your answers **short, clear and precise**!

1. Now, transform some variables using **log transformation**.

*(\* “some variables” means any of the variables in the data set, use your logic and intuition for which variables need log-transformation)*

Run a few regression models with log-transformed variables. **Pick your best model**, submit R regression output including the lm formula equation screenshot, and answer the following question:

* 1. Is this a better model as compared to the previous model for question 2? State your reason. (5 points)

1. Use your **BEST regression model** obtained in question 3 to perform “Stock Price” prediction for the following company.

|  |  |
| --- | --- |
| ***Variable Name*** | ***Actual Value*** |
| Stock Price | 4.55 |
| Total Debt | 9.5 |
| Cash | 0.3 |
| Revenues: Last yr | 61 |
| Invested Capital | 1.05 |
| Reinvestment Rate | 0.67 |
| ROE | 0.08 |
| ROC | 0.12 |
| Net Margin | 0.01 |
| Invested Capital | 15.1 |
| BV of Assets | 45.9 |
| Net Income | 0.5 |
| EBIT | 4.41 |
| EBITDA | 7.31 |
| FCFF | 0.59 |
| Cash as % of Firm Value | 0.02 |
| Cash as % of Revenues | 0 |
| Cash as % of Total Assets | 0.01 |
| Capital Expenditures | 2.2 |
| Depreciation | 2.9 |
| Trailing Revenues | 76.7 |
| Trailing Net Income | 0 |
| Intangible Assets/Total Assets | 0 |
| Fixed Assets/Total Assets | 0.38 |
| Market D/E | 1.44 |
| Market Debt to Capital | 0.59 |
| Book Debt to Capital | 0.62 |

1. What’s your point forecast for stock price (in dollars)? (10 points)
2. This company stock is currently priced at $4.55. Based on your prediction, would you buy or sell this stock? (5 points)

**Time Series forecasting (30 points)**

Download data file named “ ***GrpHW2\_ts.csv***” from eCampus. Create **monthly** time series data from it. Submit your .R file, output screenshots and visual charts for following questions:

1. Is there a trend in the data? If there is, create a new time series that’s a first order difference of the original one, and use that new time series to answer the following questions:
   1. Generate the ACF graph. What can you learn from the ACF graph? (5points)
   2. Generate the PCF graph. What can you learn from the PACF graph? (5points)
   3. Do you think there is seasonality? Of what lag? (5points)
2. Fit a suitable ARIMA model to the data using auto.arima(); submit your R screenshot on auto.arima output. ( 5 points)
3. Try some other plausible models by experimenting with the orders chosen; pick what you think is **the best model** and check the residual diagnostics. State your answer clearly why you pick this model. (5points)
4. Produce forecasts of your fitted model. Do the forecasts look reasonable? (5points)