Group Assignment 1

Write the name of all your group members at the top of the report.

Attach all pieces of the code to the respective part of the question.

Whenever you run a regression or a similar analysis, take a screenshot of the R output and report it in the respective part of the question.

In this group assignment, you are tasked to analyze data related to regression concept we learned in the class. Since we have covered regression regarding three tasks of prediction, interpretation and causality, this assignment is divided into three parts accordingly.

Part I) Prediction ----------

In this part, we want to consider the prediction power of the regression model and evaluate its performance against previous prediction models. For this part, use the "marketing_campaign.csv" dataset that we have covered previously for the task of hypothesis testing.

a) You are first tasked to run an RFM. Since this dataset does not naturally have an F or M, you should create them as we discussed in the class. To do so, use the total amount spent on all six categories of wine, fish and so on as the monetary and total number of store and web purchases as frequency. Now use **three** quintiles to do RFM and create the resulting RFM index for customers. Here, we assume that the entire population of customers is the 2240 customers we have in the dataset for economics calculations. in other words, we just calculate the profit for the sample of 2240 customers through this assignment without generalizing it to a bigger sample. For the economic calculation, we need to know the cost of sending the offer and the profit margin if the customer says Yes. The two columns in the data (CostContact and Revenue) provide us with this information: the cost of sending the offer is $3 and the profit margin of the product if the customer says Yes is $11. Yes or No responses to the offer are reported in the Response column.
   Given all this information, first, calculate and report the profit if we target everyone in the sample of 2240 customers.
b) Now instead of targeting everyone in the data, we use RFM analysis based on 3 quintiles. Calculate the profit if the company decides to run this RFM. Compare your results to the previous part and interpret it from a marketing perspective.
c) Now we run a regression of Response on R+F+M to see if we can do any better in prediction. Note here the regression is done on the actual values of data not

quintiles. Do not add any other variables to the regression. Interpret the regression coefficients. Again, make sure that you put the screenshot of the R output for the regression here. Do the reported coefficients make sense to you? Explain.

d) Now look at the predictions of the regression using the predict.lm() command. In other words, if you name your model "reg", then the command predict.lm(reg) will give you the predicted value of the outcome variable (in this case, Response) based on the regression model. After adding predicted values to the dataset, consider only customers with the highest predicted response based on the number of customers who were targeted under the RFM. In other words, look at how many customers were targeted under part b), then sort the customers based on the predicted values of the regression and select the top X number of customers where X is the number of customers that were targeted under the RFM. These customers are now going to be targeted under a regression scenario.
   Now what is the average response rate under the regression model? Compare the result to the average response rate based on RFM you found in part b).
e) Now calculate the profit under the regression case. Compare the result to the case of RFM. Do you observe an improvement? If yes, explain why you expect such improvement in the regression model over RFM?
f) Now you are tasked to make the regression result more accurate by accounting for interaction terms and nonlinearities (Only up to square terms. Do not consider higher degrees of polynomials). Do not add any new variables beyond R, F, and M. How much were you able to improve the model fit? More specifically, consider the interaction effects between the set of R, F, and M. Interpret the interaction coefficients. Are the interaction results expected based on your understanding of RFM analysis? Explain.
g) Now use the best model that you have found (using Adjusted R-squared model fit measure reported in R) in the previous part to do a different regression-based targeting. While previously we run RFM first to see how many customers need to be targeted to use as a basis of selecting customers in the regression model - refer to part d), this time, instead, we run regression on its own and decide to find an optimal number of customers to be targeted. In other way, this model will be free of any RFM analysis influence you did in part b). To do so, we define a threshold δ such that a customer is getting targeted if the predicted response using the best regression model is above that threshold. Using a function or a for loop or any other way you prefer, determine the optimal threshold δ so that the resulting profit is maximized. Based on the found optimal threshold, calculate the profit. How many customers are targeted in this case? Do we target more or fewer customers

compared to the case presented in part d)? Compare the results to the case of RFM and discuss any improvement you notice in the profit.

h) Now based on the best regression model you have found in part g), use that model to create a Gain graph based on 5 quintiles. Do the same thing for the case of RFM. Note that the RFM is still a 3-3-3 quintiles (a total of 27 indexes) and the 5 quintiles is just for the case of visualizing Gain graphs. Put the two graphs (best regression model vs. the RFM model) in your report. While you are already familiar with how to make a Gain graph based on RFM, the only difference for the case of regression is that you use model predicted responses as the basis to make 5 quintiles instead of response probabilities which we did for RFM in assignment 2. Interpret the difference you see in the two graphs. You are allowed to use Excel to plot the Gain graphs.

i) In a regular RFM, as we discussed in the class, we can just rely on three columns of R, F, and M as it is a method devoid of considering customer demographics. However, in this data set, we have the privilege of having access to customer demographics such as income, age, and so on. While RFM does not handle such demographic variables, a regression model does. Therefore, in this part, you are tasked to add a few demographic variables to the best regression model you found in part g). You are allowed to use any demographic variable in the data as long as you have a good theory of why that matters for the case of customer response. Please clearly explain your justifications before adding the demographic variables to the model. Additionally, unlike traditional RFM, we have amazing data on whether the customer has responded positively to the five previous campaigns. Therefore, use them as additional covariates inside your regression model as they might have some prediction power in predicting customer response to the most current campaign. Use adjusted R-squared as your guideline on whether adding new variables help the fit of the model or not. Once you have found the new best model, use this model once again to find the optimal threshold δ based on this new model as discussed in the part g). Then calculate the profit under the scenario that customers above the threshold are getting targeted and below threshold do not. Report on how much you are able to increase the profit with this model compared to the RFM case.

Part II) Causality ----------

In this part, we work with two different cases. In the first case, we are interested in calculating the causal impact of the showing ads to customers based on an RCT. As discussed in the class, RCTs are gold standards for establishing causality. To do so, we work with "marketing_AB.csv" dataset. This dataset records the conversion response for a sample of more than 500,000 customers. Here, a True conversion means that the customer purchased from the company and False means the customer did not. In order to see the causal impact of showing ads to customers on their conversion, the company decided to randomize customers into two subgroups of "ad" and "psa". The "ad" group are the treatment group where they see the ad in their email, while the "psa" group are not given any ads about the company, therefore they are served as the control group.

a) Using the method we discussed in the class regarding RCTs, calculate the causal impact of showing ads to customers in their probability of conversion. Hint: you also need to run a t-test.
b) Now we go one level deeper to study the impact of showing ads to customers based on the number of ads they have been exposed to before. To do that, use the "total.ads" column to create five quintiles of very low to very high exposure to prior advertisement, and for each quintile calculate the causal impact of showing ads on likelihood of purchase. Interpret the results you observe between five different quintiles. Which quintile(s) has the strongest causal impact, and which one(s) show the weakest impact? Interpret what you find from a marketing perspective.

In the second part, we are considering a different type of establishing causality where experiments are not feasible. For this part, you work with "privacy.csv" dataset.

In this case, a company is interested to see the impact of privacy regulations on online purchase behavior. The impact of having stronger privacy regulations on the adoption of e-commerce is not clear: while some customers who are predominantly privacy-conscious might increase their online presence due to a higher level of trust in handling their sensitive data by the company, companies might prefer a low regulation environment as more regulations can increase the cost of doing business for them. If that is the case, higher level of privacy regulations might decentivize companies from offering online platforms which can result in lower realized demand for online purchases. The focal company of this study

has customers across multiple states, where they are located in California, Oregon, Nevada, and Arizona. We have a panel data of 1,000 customers for two years, from beginning of 2019 to end of 2020 which span 104 weeks.

In 2020, California implemented a consumer privacy law known as CCPA (California Consumer Privacy Act) that heavily regulates the usage of customer data by a company and gives consumers a great deal of power in how to manage their online data. If you are interested, you can read more about this law in the following link:

https://securiti.ai/blog/ccpa-for-marketers/

The company is interested to see if such regulations have changed the adoption of e-commerce among their California-based customers compared to others. The company cannot do the RCT, as there is no randomization here: customers in one state (CA) have been affected in a non-random manner compared to the customers living in other states. To assess the causal impact of such policy change (implementation of CCPA) on online purchase behavior of customers, we can use a class of causal models known as **Difference-in-Difference (DID)** models. The basic idea behind this model is that we can compare the outcome among the treatment group after and before the policy change. The difference in the average outcome before the policy change and after the policy change in the treated group can show the impact of the policy. However, it is possible that there might be some other time-related shocks that affect the outcomes. For example, if customers in 2020 are more likely to be engaged in online shopping due to the COVID-19 pandemic, then simply looking at changes in online purchase behavior in California might lead to an overestimation of the effect of CCPA. To account for such time-related confounders, we introduce a control group—customers from states that did not implement CCPA, namely Oregon, Nevada, and Arizona. The key assumption in Difference-in-Differences (DID) is that in the absence of the policy change, the treatment group (California) and control group (other states) would have followed similar trends in online shopping behavior (as for example, they were also affected by the time-shock of Covid pandemic). By comparing against a control group, we make sure that our causal estimate is safe from threats of time shocks that affect everybody (both treatment and control). In other words, the causal estimate can be written as:

(Avg.outcome of Treatment group after the policy - Avg.outcome of Treatment group before the policy)

-

(Avg.outcome of Control group after the policy - Avg.outcome of Control group before the policy)

In other words, we are taking two differences, and hence the name Difference-in-Difference. Using panel data and the notion of interaction effect, we can extend this DiD notion to the regression setting:

$$Y_{it} = \alpha + \beta \times T_i + \gamma \times after_t + \eta \times T_i \times after_t + \epsilon_{it} \qquad \text{Eq.1}$$

Here Y_{it} is the outcome variable for customer i at week t, α is the intercept of the regression model, T_i is a dummy variable that shows whether the customer i belongs to the treatment group ($T_i = 1$) or the control group ($T_i = 0$). Variable $after_t$ is a dummy variable that shows whether week t is after the policy change ($after_t = 1$) or before the policy change ($after_t = 0$). In this setting, η is the causal estimate of the policy on the outcome. In order to see this, we can simply consider the following 2*2 table based on dummies for treatment and control, as well as dummy for before and after the policy change.

	Before the policy change ($after_t = 0$)	After the policy change ($after_t = 1$)
Control group ($T_i = 0$)	$\bar{Y}_{it} = \alpha$	$\bar{Y}_{it} = \alpha + \gamma$
Treatment group ($T_i = 1$)	$\bar{Y}_{it} = \alpha + \beta$	$\bar{Y}_{it} = \alpha + \beta + \gamma + \eta$

Here $\bar{Y}_{it}$ is the average outcome. First row and first column gives us the estimate of α. Now if we take the difference in the first row, we get $(\alpha + \gamma)$ - $\alpha = \gamma$. Now if we take the difference in the second row, we get $(\alpha + \beta + \gamma + \eta)$ - $(\alpha + \beta) = \gamma + \eta$. Now if we take the difference between these calculated differences we get to $(\gamma + \eta)$ - $\gamma = \eta$. So now we successfully linked the following measure:

(Avg.outcome of Treatment group after the policy - Avg.outcome of Treatment group before the policy)

-

(Avg.outcome of Control group after the policy - Avg.outcome of Control group before the policy)

To the estimate of η. In other words, to calculate the causal impact of the policy change we just run Eq.1 regression and interpret the coefficient of interaction effect $T_i \times after_t$ as the causal impact of the policy change on the outcome.

Now you are tasked to consider the impact of the CCPA policy on online purchase behavior of customers. In this setting, California is the treatment group and all other states are control group, all the weeks in 2019 (first 52 weeks) are before the policy change and weeks 53-104 are after the policy change. The dataset uses the column “treatment” to show treatment dummy status and “after” to show the after status dummy.

c) Using the online ratio variable - which shows the ratio of online purchases to the total number of purchases happened in the week for customer, run the DID

regression. Report the coefficient table and interpret the causal impact of the policy change on online purchase behavior (i.e., the coefficient of η). Does this result make sense to you, explain.

d) After finding the causal impact of the policy change on the online purchase behavior of customers, the company is looking to study such impact deeper through considering customer heterogeneity. The dataset has three demographic variables on customer income, age, and political affiliation. Add these variables to the regression and interact them with $T_i \times after_t$ in your model. Doing so will provide us with valuable insight on whether there is heterogeneity in the impact of the policy change on online purchase behavior based on different demographic groups. For example, maybe younger patients care more about privacy and thus the impact of implementing a privacy law can increase their online presence more than older customers. Interpret the causal impact with respect to income, age, and political affiliation. Provide a marketing insight into the results you found.

Part III) Interpretation

For this part, you work with the "bank.full.csv" dataset. This data is about a bank who ran a campaign for more than 45000 of their customers and the marketing department is interested in interpreting the results of the campaign so that they can run better campaigns in future. You can see the description of each variable in the dataset at the end of this document. Our main goal is to investigate the impact of two important marketing measures in running campaigns; namely the duration of the phone call to the customer and the frequency of the contacts the bank has made for the specific customer. In the dataset, these are recorded as "duration" and "campaign" columns. Our objective is to study the impact of these two measures on the response to the campaign (recorded as yes or no). Define y = yes as 1 and y = no as 0 for the case of this assignment.

A) Run a regression model to study the impact of duration and campaign on y. Check for interactions between these two variable of interest. Make the model more accurate by accounting for more factors as control variables. Report regression results.

B) Interpret the coefficients you obtained for duration, campaign, and the interaction term between duration and campaign. In doing so also explain how much y changes if duration or campaign go up by 1 unit. What does this tell us

C) What is the impact of candidate's job title on the chance of accepting the campaign offer? Rank different job titles in terms of their likelihood to respond to the

campaign. **Hint:** you need to compare regression coefficients to each other and determine if the difference between coefficients is significant.

D) Can we claim causality in the relationship of either duration or campaign on the response (y)? If no, name at least one confound that is not in the current dataset. Clearly explain your proposed confound by explaining why the confound is correlated with duration/campaign and why it can affect response to the campaign. Explain if there is any reasonable way for the bank to obtain data on that confound.

E) Propose an RCT to establish causality for each case of duration and campaign. Explain how the bank should run the proposed RCTs.

F) The marketing team is now exploring changes to the structure of the campaign. The team believe that there might be missed opportunities in increasing the response rate to the campaign. As a result, they are now considering two alternatives to the campaign. In the first alternative, the bank will look at the outcome of the previous campaign (recorded in the data as "poutcome". If the value of the variable is "failure" for a customer, the company spends less resources on them by reducing the number of calls (campaign variable) by 30%. If the value of the variable is anything but failure, the company does not change course (no reduction in number of calls). In the second alternative, the company is looking to spend more time with customers who have responded positively to the previous campaign (i.e., poutcome = "success") through increasing the duration of the phone call by 20%. If the value of the variable is anything but success, the company does not change course (no increase in duration of calls). Your task is to help the marketing team to decide which one of these alternatives yield a better result for the bank in terms of increasing the likelihood of getting better response (greater y). **Hint:** run the regression model. Then and create two new datasets based on the proposed changes. Using predict.lm() command get predictions for y and see under which alternative the difference between average y before any change and after the change is greater. This is known as counterfactual analysis.

Description of Bank-full dataset

Age Age of customer
Job Job of customer
Martial Martial status of customer
Education Customer education level
Default Has credit in default?
Housing If costumer has housing loan
Loan Has Personal Loan
Balance Customer's individual balance
Contact Communication type
Month Last contact month of year
Day Last contact day of the week
Duration Last contact duration, in seconds
Campaign Number of contacts performed during this campaign and for this client
Pdays Number of days that passed by after the client was last contacted from a previous campaign
Previous Number of contacts performed before this campaign and for this client
Poutcome outcome of the previous marketing campaign
Y has the client subscribed a term deposit