MSSP PORTFOLIO

Xinci Chen

Boston University
Department of Mathematics and Statistics

Table of Contents

External Partner Project (Fall 2019)

Introduction

The external partner project collaborated with a Boston-based investment firm is a two-semester project. This firm is ranked in the top 10 US investment firms by assets. It is financial service company that help people with different investment products, for example, mutual funds. Since the clients' interest and data provided are differ within two semesters. This part will be only focus on the Fall semester.

Mutual funds pool money from the investing public and invest in securities such as stocks, bonds, money market instruments, and other assets. The value of the mutual fund company depends on the performance of the securities. As an investment company managing a large family of mutual funds, being able to project future performance for mutual funds accurately is very important to our project partner. A standard method to project future performance for a fund is to use an index, whose performance correlated to the target fund's performance, to project the net asset value (NAV) of the fund. Therefore, the project objective is to find a list of indices that can project the net asset value of the mutual fund also outperform the benchmark index.

Data and Methods

In the fall semester, we will not be in touch with the client's internal Data. The response variable is NAV of each fund is collected in Yahoo Finance from year 2014 to 2018. The benchmark information is listed in the company website. Besides the NAV data, we also collected a lot of different indices NAV data online. Due to the difference of each fund, the indices will be slightly different. The more general indices we chose are S&P 500, Nasdaq composite and Russell 2000.

The performance of funds and indices is directly reflected in their close price (NAV for funds). The way we approach the fit of the funds and indices is to model on their log return. When using log return we are essentially assuming that price of assets has log normal distribution, therefore the return is normally distributed, which fits the assumption of linear regression. The baseline of fit is set up by regressing the log return of the fund on the log return of the index. Then with the indices data for each fund, we examine the fit of individual index using linear regression. Any individual index that can outperform the benchmark index in terms of fit would be recorded. In this way, we check the fit of individual index, which can later be compared to the fit of composite index. Next, we use lasso regression to select indices and build a composite index. Lasso (least absolute shrinkage and selection operator) is a regression analysis method that can select model variables, which is achieved by adding a L1-norm as the penalty term to the log-likelihood function. This penalty term works as a constraint on the sum of the absolute value of the magnitude of coefficients, which lead to a regression model with fewer coefficients as indices whose coefficients get close to zero would be eliminated.

Result and Discussion

Using the Lasso Regression, each fund has a list of composite indices to project the NAV. Most of the result outperform the benchmark index. But there are cases that the benchmark has a higher accuracy rate projecting the NAV. The combination of the result of 15 mutual funds are some general indices which is not surprising. The more fund result we put together, we would expect more general indices appears rather than a specific index to some funds.

External Partner Project (Spring 2020)

Introduction

In Spring 2020, Our goal is to projecting daily net flows for year 2018 (given the dataset from 2014 to 2017) for 15 mutual funds. The technical requirement for this project is to build a fully automated and scalable work which can be reproducible.

Data and Methods

The data was provided by our client internally, which is also highly secured data. We have received the data of 15 mutual funds. Each mutual fund has three datasets all from year 2014 to 2018. The datasets are total assets data, net asset value data and flow data with inflow and out flow. Each flow also has its own transaction type, transaction category and account type. And we also have some market indices from the last semester result.

The flow data comes with a lot of noise. We have built a universal error and outlier handling function to help us to adjust potential error. Different method was conducted during this semester. Since overall the net flows is a time series data, we focus more on the time series method including ARIMA, ARCH/GARCH, Prophet, VAR. Non time series method we used for prediction is random forest. Our baseline model for evaluation the performance of prediction is the average of past three-year net flows. We have built universal prediction average asset function based on our baseline model to help us selecting best models among all. All these functions allow a uniformed automated pipeline that can be easily applied and scaled for future use. ARIMA and ARCH/GARCH have the better performance comparing with other models. ARIMA and ARCH/GARCH method are both in the time series forecasting field. ARIMA is used for modeling the level of the series. ARCH/GARCH are used for modeling clustered volatility of the financial series data. It takes in to account that the complex volatility of financial data varies over time and estimate the mean. Since more volatile means more uncertainty in the forecasting. GARCH is important if we suspect volatility changes over time.

Result and Discussion

Even though ARIMA and GARCH model have a better performance overall. They still did not outperform the baseline model. Therefore, a simple ensemble model brought up by the team. It is the average of the prediction of ARIMA, GARCH and baseline net flow. The result of ensemble is improved. It outperforms the baseline model.

Variety of Capitalism Project

Introduction

Our client is Songhyun Park, a Ph.D. student from Department of Political Science. She has been conducting research on clarifying the identification of market economies on both the state and national level of the United States. According to the Varieties of Capitalism theory, United States is defined as a liberal market economy, and one characteristic of this market type is that the union density is positively related to the vocational education. However, during her research process, she found a negative relationship between union density and vocational education, which conflicts with the Varieties of Capitalism theory. Her research question then is whether racial diversity can explain the negative relationship between union density and vocational education. Two types of analysis were used for this project based on the research question. One is to validate the negative relationship between the union density and vocational education using regression modeling and another is to use mediation analysis to determine whether racial diversity mediates the relationship between union density and vocational education.

To be clarified, we received two datasets from the client and did two analysis according to the data. We have some concerns on the collection and structure of the first dataset we received. We did some exploratory data analysis and replicate her initial model with this dataset. After presenting our concerns and result to the client, she sent us another dataset which she collected under our suggestion. Another full analysis was conducted on this dataset and results were presented to the client. Therefore, the following will be consisted of two parts. The first part is about the first dataset, the second part is about the second dataset.

Part I

Data and Methods

The first dataset contains 31 variables for 50 states in the US and Washington D.C. The dependent variable is the number of high school students in the vocational education program, which represents the vocational education. And the independent variables include the predictor and confounding variables. The predictor is the union density, and the confounding variables are proportion of White, Black and Hispanic people, GDP, urbanity, population, export, cost of living, unemployment rate and Gini index.

There are several issues with the dataset that may have an implication on the analysis we conduct. First, data are sourced from different years, from 2016 to 2018. Second, the number of high school students in vocational education programs may not be representable for vocational education.

The exploratory data analysis does not suggest a distinctive pattern of the negative relationship between vocational education and union density, therefore we replicated the multiple linear regression used by the client initially to confirm the model results. We were able to replicate her results. But we find out that the original linear model does not align with the assumptions of the linear regression. Therefore, we transform the data so that assumptions are not violated. Because there are other factors that can impact the relationship between vocational education and union density, a partial correlation analysis is also implemented to better

investigate the relationship between the variables. The partial correlation is a measure of the strength and direction of the relationship between two continuous variables, while controlling for the effect of other predictors in the regression model.

Result and Discussion

First, there is no significant relationship between union density and vocational education from linear regression after transforming the data. Secondly, the racial diversity changes the results of linear model to some extent. However, for the question of whether the racial diversity explains the relationship, no conclusion can be made. Therefore, in the second part, we conduct mediation analysis to further explore whether racial diversity mediates the relationship between union density and vocational education.

Part II

Data and Methods

After the client takes our suggestion and put more thoughts into the confounders that need to be considered, the new dataset consists of 8 variables for 50 states in the US and Washington D.C. The dependent variable is the Number of vocational education concentrators, and the independent variables include the predictor and other confounding variables. The predictor is the Urbanity, and the confounding variables are unemployment rate, GDP Per Capita, Union, Racial(black/Hispanic/white), Total population, Student population.

The initial Method is the same as the Part I, we used multiple linear regression and transformation on the variables. Besides the linear model, to investigate whether racial diversity actually mediates the relationship between union density and vocational education, we conducted a statistical mediation analysis. A mediation model proposes that the independent variable influences the mediator variable, which in turn influences the dependent variable. In this project, the union density is the independent variable which serves as the treatment and vocational education is the dependent variable, which acts as the outcome. Racial diversity is the mediator. Thus, here, the mediation analysis we conduct explores whether union density affects racial diversity first, and racial diversity then affects vocational education. Because mediation analysis is a causal inference model, we assume that there are no unmeasured confounding variables in the model.

Result and Discussion

The linear regression shows that union density has a slightly negative relationship with vocational education. However, these models are built based on the assumptions that we assume there are no other unmeasured confounding variables. Furthermore, through the mediation analyses, we see that racial diversity has no mediation effect on the negative relationship between union density and vocational education. Therefore, we cannot conclude that racial diversity explains the negative relationship between union density and vocational education.

Coral Host and Symbiont Project

Introduction

Our client is James Fifer, a Ph.D. student from the Department of Biology. He is interested in investigating the difference in gene expressions for coral host and symbiotic algae under different flow and heat conditions. He conducted two separate Principal Components Analysis (PCA) on the coral host and symbiont algae. And the two different PCAs were being compared and interpreted using their PCA scores. This approach raises our concerns because the PCA performed on different data frames separately are not necessarily comparable. Therefore, our goal is to find a reasonable method to help our client to compare two gene expression datasets (coral host and symbiont algae) in a reduced dimension space.

Data and Methods

Multiple co-inertia analysis (MCIA) were considered a reasonable way for our research questions. MCIA is a way which can project k datasets onto the same space for comparison. Since our client experienced more on dealing with marine biology data than our team and have a standard statistical background. Instead of perform a full MCIA on our client's data, we decide to use a simple but similar data as an example to show and guide our client how implement MCIA in R.

The toy datasets are a subset of microarray gene expression of the NCI-60 cell lines data stored in R, which contain list of 3 data frames, the mRNA data frame which contains 12895 rows represent gene expression measurement and 21 columns represents cell lines, the miRNA data frame which contains 7016 rows represent gene expression measurement and 21 columns represents cell lines and the proteomics data frame which contains 537 rows represent gene expression measurement and 21 columns represents cell lines. MCIA was applied to analyze mRNA, miRNA, and proteomics expression profiles of melanoma, leukemia and CNS cell lines. MCIA links the individuals in the different datasets, and thus the columns will be linked between the multiple datasets.

Result and Discussion

The sample space plot and the eigenvalue plot generated from MCIA method are the most important plots to understand the relationships between our sample data or between the host and symbiont. The sample space plot mainly explores how the clustering of different data frames are associated with gene expression from host and symbiont. The eigenvalue plot indicates how much variance is explained in the datasets by the corresponding eigenvectors. The algorithm only finds the best subdivision of the original feature space based on the written mathematic equations. It does not consider the meaning of the feature or any other information about the feature.

Orangutan Seed Dispersal Project

Introduction

Our client is Andrea Blackburn, a Ph.D. student from Department of Anthropology. The purpose of the project is to investigate whether the fruit availability or age-sex class of orangutans have a greater impact on the total number of seeds found in the orangutans' seed dispersal behavior.

Data and Methods

The fruit availability is measured by using the percentage of mature and ripe trees per month (per_MR). per_MR is calculated by dividing the total number of trees with mature or ripe fruits by the total number of trees. And there are four types of age-sex classes: females, flanged males, unflanged males and infants. The response variable is total number of seeds found in the orangutan's feces. However, we believe that we should take the weight of the feces samples into account because if the sample is larger, there could be more seeds. The total number of seeds per unit sample weight is calculated by dividing the number of seeds by the sample weight of the feces.

First, we fit a regular Poisson model because the response variable, the total number of seeds, is a count variable. However, we suspect there may be an inflated number of zeros based on exploratory data analysis (EDA). Thus, we fit a regular Poisson model on the data to evaluate the number of zeros and the dispersion to determine the appropriate model to use. And also include orangutans and month as a random effect based on the findings of EDA. The distribution of the residuals is not perfectly normal, it is probably due to an excessive number of zeros presented in the data. Thus, we test the zero inflation and over dispersion. The dispersion test checks whether residual variance is larger or smaller than expected under the fitted model, and the zero-inflation test compares the distribution of expected zeros in the data against the observed zeros. From the diagnostic plots for the Poisson model suggest that the data does contain more zeros than expected and dispersion is as expected. So, we use the zero inflated Poisson model instead of a zero inflated negative binomial model. A zero-inflated Poisson model is used to model count data that have an excess of zero counts. Also, per_MR variable ranges from 0.03 to 0.1, while the age_sex variable is binary. Therefore, we standardized the per_MR variable. The main analysis method used is a Zero- Inflation Poisson model where the per_MR variable is standardized.

Result and Discussion

We found that the percentage of mature and ripe fruit is more significant than the age-sex class in the data provided. The age-sex class is inadequate to draw any confirmatory conclusions based on current data. The limitation here is that we are performing available case analysis, which assumes missingness is completely at random, which is most likely not the case. In particular, we saw that the coefficients for the infant and unflanged male orangutans are not statistically significant. This suggests that only the flanged male orangutan is significantly different from the female orangutan. While this may be true, there are issues with

the coefficient estimates for the age-sex class besides females. The data itself is imbalanced in at least two critical dimensions: age-sex class and month. Overall, the observations are predominantly female orangutans, and the number of records varies quite a lot from month to month. Because there are missing data for some months for flanged males, infants, and unflanged males, it creates a systematic correlation between the age-sex class and the month variable, which will make the coefficient estimates for both variables unreliable. Another limitation comes from the potential confounding covariates. High sample weight does not necessarily associate with a higher seed count. In the data, June has the highest overall sample weight, but the seed count is among the lowest throughout the year. This could be an artifact of little observations, but it could also indicate a change in the orangutan's diet that resulted in a smaller seed count. These effects have not been accounted for by the model, and they might have a significant impact on the actual association.

BU Athletics Marketing Project

Introduction

Our client is Brendan Sullivan, a senior associate director of athletics from BU Athletics Marketing and Communication Team. The purpose of this project is to investigate what promotions drive attendance in order to best prioritize the marketing budget in regards to Boston University's 6 ticketing sports.

This project is a collaboration with another consulting team. And since two datasets were received from our client. Each of the team is responsible for one dataset. Our team received the season ticket holder survey data. And due to the timeline of project, the goal for the survey portion is to provide the exploratory data analysis (EDA) in regards to the promotions being done at the games and to provide recommendations on building effective future surveys.

Data and Methods

The survey data for the past two years (2017-2018 and 2018-2019) is provided. We have data for all six sports for 2017-2018, and we have data only for basketball and women's ice hockey for 2018-2019. Since we do not have a lot of observations from the survey data in general, and the survey data for only two sports were provided in 2018 - 2019, we decided to combine all the data together to have a broader view of the survey data. When combined, the dataset contains a total of 68 responses.

Since a lot of the EDA on the categorical variables was already done by our client, we decided to focus more on the text box questions, especially the promotion ones. We explore the frequency of words in the answers by subdividing the answer's text into bigram/trigram (consecutive sequences of 2/3 words). Whether a bigram or trigram is used depended on whether a bigram or trigram provided more contextual information regarding the answers. Word clouds (one word) of the text box questions are also provided here because some information may be missed when we group the words by in two/three.

Result and Discussion

The EDA shows us what season ticket holder preference on promotions. It seems that the respondents enjoy the season ticket holder appreciation nights the most, as well as club sports night, trivia night, and chili cook-off night.

There are three major concerns regarding the promotion question in survey. First it is a required text box question, which leads to difficulties when trying to make evaluations. Secondly, the question required the respondent to name the promotion, which in many cases, the respondent could only remember some features of the promotion, like "free food". This will introduce bias to the answers because respondents may tend to name promotions that occurred closer towards the end of the season or promotions that are more unique since those are more memorable. Lastly, because different sports have different promotions, the imbalance of the number of respondents between sports could lead to inaccuracies in the frequencies of survey

responses. And for the survey in general, the response rate seems fairly low and so the sample of survey responders may not be representative of even the season ticket holder population. However, we do not have any information about the total number of surveys sent, so we are not sure how low the response rate actually is. The current survey can also only answer questions regarding what, specifically, season ticket holders like/dislike about their experiences of the events in which they attended and can remember. However, it does not answer the question of what aspects, in particular, may attract season ticket holders to actually attend games/events. Furthermore, since the participants are season ticket holders, these responses may be biased and their focus and experiences may not be able to be generalized to other general fans.

Based on the concerns, we suggest that change the format into a checkbox question and to provide a list of promotions for the promotion question. Such a format can simulate more robust answers from the respondents and can provide more meaningful insights. And a survey sent to the general public will be better in order to get a better idea of what attracts an audience.