Regression Analysis: Project Assignment
Year 3, Institute of Technology of Cambodia

Department of Applied Mathematics and Statistics

Guideline

The followings are important information about the projects:

- This project is worth 30% of your overall grade.
- You can do this project by using any statistical software you are most comfortable with namely Ms Excel, Python, R programming etc.
- The report **must not** be longer than 10 pages excluding appendix and references.
- The report **must** be written by either latex, overleaf or R markdown.
- All statistical tests are tested at a 5% level of significance or otherwise specified.

Instruction

You can choose your most preferable topic/research question from the **Project** section (last section) in this file. To own a project, please fill this form your group number in the same row as the project you want. For choosing the topic, we are using a first-come-first-serve policy meaning that you are not allowed to replace an existing group number with your group number if you are late. If found out, your group will get 0 for this project assignment.

One topic/research question is allowed for **only one** group to do. You can also propose your original project if you do not wish to do any of the available projects.

For groups that choose projects with code starting with ITC (e.g. ITC002), please think of how large the sample size and what are relevant variables you need to answer the research question.

Timeline

22 Feb	
01 Feb	
08 Mar	Project Launch
15 Mar	
22 Mar	First Presentation: EDA (5%)
29 Mar	
05 Apr	Second Presentation: Data Analysis (5%)
12 Apr	
19 Apr	
26 Apr	Project Report Deadline (10%) and Final Project Presentation (10%)
03 May	
10 May	

Project

Code: RA001

Predicting student performance using multiple regression analysis. You can use data from the Student Performance Dataset: https://archive.ics.uci.edu/ml/datasets/student+performance

Code: RA002

Examining the factors that influence car prices using multiple regression analysis. You can use data from the Car Price Prediction Dataset: https://www.kaggle.com/hellbuoy/car-price-prediction

Code: RA003

Predicting the success of a movie using regression analysis. You can use data from the MovieLens Dataset: https://grouplens.org/datasets/movielens/

Code: RA004

Investigating the factors that influence housing prices using multiple linear regression. You can use data from the House Prices dataset: https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Code: RA005

Analyzing the factors that contribute to employee turnover using logistic regression. You can use data from the IBM HR Analytics Employee Attrition Performance dataset: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

Code: RA006

Predicting customer churn using logistic regression. You can use data from the Telco Customer Churn dataset: https://www.kaggle.com/blastchar/telco-customer-churn

Code: RA007

Analyzing the factors that affect customer satisfaction in the airline industry using ordinal logistic regression. You can use data from the Airline Passenger Satisfaction dataset: https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction

Code: RA008

Examining the relationship between social media engagement and sales using linear regression. You can use data from the Social Media Analytics dataset: https://www.kaggle.com/c/avito-context-ad-clicks/data

Code: RA009, Linear Regression Project Idea for Stock Price Prediction

The first thing that comes to mind at the mention of finance is strangely most likely 'stocks'! This is perhaps because companies and individuals alike stand to invest and make money in stock markets. Predicting stock, although enigmatic, can, therefore, be a great area to explore.

Stock price prediction can be of great importance to investment brokers and (potential) investors alike, and a good forecast prediction can mean actual financial payoffs. The publicly available Kaggle dataset of the Tesla Stock Data from 2010 to 2020 can be used to implement this project. Maybe you could even consider gathering more data from the source of the Tesla Stock dataset. A multiple linear regression model can be used for the implementation of this regression analysis project idea. You could also try the modifications like Lasso or Ridge regression to understand the impact of regularisation on model performance.

Code: RA010, Logistic Regression Project Idea for Loan Default Prediction

Sanctioning a loan is an essential decision for any lending institution. A bank loses out on potential income by rejecting a loan to an individual or a company. At the same time, granting loans where lending risks exceed the returns could result in heavy losses. This is why banks stand much to gain from relying on good loan default prediction models based on actual statistics. You can implement this logistic regression project on the SBA Loans Case Dataset, including historical data from 1987 through 2014.

The use case can be solved using a two-step approach as some of the winning solutions for a similar challenge on Kaggle. The first step, in this case, would be to predict the defaulters using a classification technique like logistic regression. Subsequently, you can predict the loss using gradient boosting regression, support vector regression, or even ordinary multiple linear regression.

NOTE: Although logistic regression is used for classification problems, it is very much a regression model at its core. It transforms into a classification technique only when the decision threshold is considered.

Code: RA011, Medical disease diagnosis

With the overload on the healthcare system (especially in the present scenario), it is often difficult for healthcare services to triage the cases they are presented with. This often results in patients not receiving the care they require, misdiagnosis due to human error, and even complications owing to delayed treatment. The use of machine learning models, while not spared from its share of skepticism, has recently taken over as a great way to serve as a preliminary filter to allow early diagnosis in the healthcare sector.

With improved models and a deeper and more widespread understanding of the machine learning domain, most people have come to accept that a human doctor will very likely miss any case missed by a well-developed model. Human intervention in diagnosis cannot be taken away entirely (at least not just yet); the extra aid that machine learning offers could help.

For this logistic regression project, you could consider undertaking the task of building a model to diagnose PCOS using the Polycystic Ovary Syndrome (PCOS) Dataset. To accomplish this task, you could use Logistic Regression. To stay true to our purpose of achieving triage, make sure you output the probability of disease being present rather than the classification, as is usually the case when the techniques mentioned above are used. It might be essential to select the correct features to build your model. Checking the correlation between the features and visualizing the features could help in this effort.

Code: RA012, Regression Project for Predicting Diseases

Predicting disease is different from the previous project, i.e., disease diagnosis, in that the prediction is not made based on tests screening for the disease but rather on features like family history, behaviors and habits, environmental factors, and even genetic markers. Suppose you consider that mundane data like proximity to highways and food habits can be used to predict the likelihood of disease. In that case, the idea itself is revolutionary. While hospitals and insurance providers might gain from using such predictive models, individuals are definitely the most significant beneficiaries as a mere change of habits or residence could help them remain in good health.

You could undertake this exercise using the publicly available Cervical Cancer Risk Classification Dataset. For this problem, you could use the Support Vector Machine. However, ensure a probabilistic interpretation of its results to stay true to our purpose of accomplishing regression.

Code: RA013, Linear Regression Project for Medical Insurance Forecast

Insurance companies need to set the insurance premiums following the population trends despite having limited information about the insured population if they have to put themselves in a position to make profits. This makes it necessary to estimate the average medical care expenses based on trends in the population segments, such as smokers, drivers, etc.

To implement this regression project example, you can use the Medical Cost Personal Datasets available on Kaggle. The aim here will be to predict the medical costs billed by health insurance on an individual given some or all of the independent variables of the dataset. Since the cost to be predicted is a continuous variable, it is pretty natural that regression is to be applied in its truest form (i.e., without the decision boundary as in regression-based classification). Therefore, you

could choose to implement polynomial, multiple linear regression, or even Elastic Net Regression. Exploratory data analysis can be an essential step (even in this case despite the limited features). You will observe patterns, like the decreased tendency to smoke among those having children, helping you achieve reasonable feature selection and simpler models.

Code: RA014, Multivariate Regression Project Idea for Movie Rating and Revenue Prediction

With production costs sometimes going over $100 million, films can be a significant investment. Yet predicting whether a film will most definitely be a success has more than just the money that goes into its making. Therefore, using box office prediction models can be a safe and logical way for investors to choose films to invest in and, more importantly, a great way for movie creators to maximize the odds of their film" success.

One can implement this project on the TMDB 5000 Movie Dataset. Alternatively, you could build your custom dataset with The Movie Database API from where this dataset was originally created. As the objective of this regression project is to predict the revenue and rating, you will need to use ML regression models capable of handling more than one dependent variable, such as the multivariate regression model

Code: RA015, Linear Regression Project for Sales forecasting

For the smooth running of businesses, the operating expenses should be matched by the sales and exceed them to make profits- which, if we're being blunt, is the only objective of running businesses. Sales forecasting, another example of time series analysis that deals with time series-based data, can be a valuable tool to evaluate sales performance and, consequently, identify shortcomings in current operating procedures and the problem areas.

You can implement this regression project with the Superstore Sales Dataset, which comprises four years' worth of retail data from a global superstore. You can try to implement this project with an advanced regression technique like the Random Forest Regressor. You could also attempt to compare this model's performance to the results obtained using ordinary linear regression or even multiple linear regression.

Code: RA016, Linear Regression Project for Pricing Strategy

Pricing often determines the future of a product because it influences a product's appeal to customers and the profit margin on the product. A combination of factors such as competition, market situation, brand value, and target customer group need to be considered before deciding on a price. A machine learning model which estimates the price based on all of these factors can, consequently, be an invaluable resource, at the very least for the initial estimation stages.

Use the electronic product prices Kaggle dataset consisting of pricing details of over 15,000 electronic products to analyze the pricing strategy and subsequently estimate the same based on training data. You can either use multiple linear regression or one of the more advanced regression techniques to solve this problem.

Code: RA017, Linear Regression Project to Analyse Social Media Marketing Data

Social media marketing is an essential means of digital marketing. In comparison to traditional marketing, the ready availability of information regarding the outreach, views, and acceptance of digital marketing campaigns makes it almost wasteful not to fine-tune marketing strategies by tapping into and analyzing the information. To get a hands-on feel of analyzing marketing data, you can use the marketing dataset that comes with the datarium package in R. Compare the performances of simple linear and multiple linear regression to arrive at a model with a reasonable level of accuracy for this problem.

Code: RA018, Regression Project for Marketing Outcome Prediction

Most companies invest in various marketing mediums such as television advertisements, social media and influencers, and radio promotions to attract customers and advertise products. Further, with the advent of digital marketing and the ubiquitous availability of information, data analysis is becoming

increasingly important. Use the advertising and sales dataset available on Kaggle to predict the sales resulting from expenditure towards various marketing mediums. Use regression to achieve this objective, draw insights on which marketing mediums have the highest impact, and attempt to implement cross-validation in your solution to avoid over or under-fitting.

Code: RA019, Customer Ad Clicks

Data regarding the number of customers engaged by an advertisement or the number of ad clicks flows in continuously and changes in real-time. Therefore, the model for such data also needs to change as the data flows in for more accurate predictions. You can implement this logistic regression project using the Predicting customer ad clicks dataset and build a Bayesian Logistic Regression model since this model will be well suited to incorporate the real-time nature of the data expected. Make sure you focus on obtaining the probability of ad clicks rather than the classification. Knowing the likelihood will help identify potential problems and design the advertising strategies better.

Code: ITC001

Research Question: Do Bac II grades affect the performance of ITC students in mathematical courses like Calculus, Linear Algebra and Statistics in foundation years after controlling other relevant variables? Dataset to be obtained from a survey.

Code: ITC002

Research Question: Do the provincial students in ITC perform differently in mathematical courses like Calculus, Linear Algebra and Statistics than students from the capital, Phnom Penh, after controlling other relevant variables? Students from which province perform the best? Dataset to be obtained from a survey.

Code: ITC003

Research Question: Are there any differences in the performance of single students and students who are in a relationship in mathematical courses like Calculus, Linear Algebra and Statistics? Dataset to be obtained from a survey.

Code: ITC004

Research Question: Are there any differences in the performance of ITC students in mathematics courses like Calculus, Linear Algebra, and Statistics based on their type of shelters in Phnom Penh e.g. family house, rented room, dormitory, Buddhist pagoda, and Christian church? Dataset to be obtained from a survey.