Debasis Chatterjee
Syracuse iSchool
Applied Data Science
SUID: 233176962
Resume
@LinkedIn

PORTFOLIO

REPOSITORY

GitHub Link

ABSTRACT

This paper is the Graduation Portfolio Milestone for my master's degree of Applied Data Science at Syracuse University. This also includes summary of my growth and reflection regarding my experience in the last 2 years. It focusses on my knowledge about data science and machine learning through the major academic projects I did in the previous coursework. Within the span of 2-year time, the program exposed me to various courses to prepare me to take on a variety of problems - text mining, time series data analysis, information visualization, data warehousing and so on. I cherry picked academic projects that showcases different areas of expertise and skills sets, including python, spark, SQL, and R. You can follow my journey of knowledge and critical thinking regarding data science, and the technical skills that are mandatory to succeed in this field have progressed along the way. My portfolio milestone will cover my learning reflection from the 5 projects I picked, my latest resume, and the github repository with the related files and codes.

SALIENT POINTS ON HOW DATA SCIENCE MASTER PROGRAM HELPS

This paper is basically a summary of what specialization I have earned during the 2 years in the master's program of Applied Data Science - the knowledge and critical

thinking regarding data science, and the technical skills that are required to succeed in this field.

Data science, without a doubt, has been the hottest buzzword in the past decade, it further creates the most demanding job opportunities in the current society. It makes businesses strategize their marketing strategies more targeted and effective, helps banks and associated authorities identify and even prevent any potential frauds, and supports researchers to spot brand-new patterns or characteristics that might bring about a medical breakthrough. Data science, by its name, is a discipline about data. However, this topic involves a wide range of aspects, including data analysis, text mining, business intelligence, deep learning, machine learning and many more. Each area of study can further ne fragmented to a variety of tasks, such as data collection, exploration analysis, data transformation and preprocess, model development, derived business insights and advanced actions. Data science is a vast and complex topic that cannot be capsuled in this paper, but I picked six academic projects that covers different areas of expertise - text mining, time series big data analysis, information visualization, data warehousing and data mining. I would introduce them in chronological order so that you can see my knowledge and skills progress.

ACADEMIC PROJECTS

Movie Recommender

The main goal of this project was this project is to develop a collaborative filtering recommender system for movies. If two users share the same interests in the past, e.g. they liked the same book or the same movie, they will also have similar tastes in the future. .
The collaborative filtering approach considers only user preferences and does not take into account the features or contents of the items (books or movies) being recommended. In order to recommend movies a large set of users preferences, towards the movies from a publicly available movie rating dataset, are used.

Technologies: R, ggplot, Illustrator, Shinyapps

Techniques: data mining, descriptive analytics, grouping and aggregation, plotting, data transformation, illustration, color theory, k-fold cross-validation, K-Mean Clustering, Item-Based Collaborative Filtering, User Based CollaborativeFiltering

Ethical consideration: Move would be recommended to user based on his item interest of based on past habits or from a large set of movie rating dataset where same interest considered into account.

Project document: Movie Recommender

Yelp dataset Analysis

This project aims to solve a financial investment problem. Given an investment opportunity in real estate, what three zip codes would yield the best return. This is obviously a prediction problem and because the data is median home values per zip code reported each month, it's also a time-series problem. Time-series analysis offers interesting statistical challenges such as transforming data to achieve stationarity and testing stationarity using the Dicky-Fuller test as well as testing for auto-correlation. The standard forecasting model used for time-series, ARIMA, was not used in this project in favor of a new library by Facebook called Prophet.

Technologies: R, ggplot, python, jupyter notebook, JSPN, PANDA, Numpy, NLP LDA, NLTK, pyplot, seaborn, basemap

Techniques: data mining, plotting, data transformation, predictive analytics, machine learning, time-series, test statistics

The analysis, including modeling, covered the following areas:
• Text mining to derive review sentiment,authenticity,sentiment analysis
• Topic modeling of reviews to segment reviews
• Review count, rating forecasting and seasonality analysis
• Photo analysis for photo segmentation
• Geo visualization for business selection

Observations were derived:
• Seasonality identification from forecasting
• Conflict between star ratings and sentiment
• Conflicts between star ratings and authenticity

Geographic analysis on business groups:
• Social Media strategy can be enhanced by reviewing these results
• Business cycles can be identified using these results
• Comparative behavior can be derived based on topic modeling

Models/Methods are used:
• verify authenticity of reviews - Naïve Bayes
• predict sentiment by review - SVM
• Topic modeling - LDA (Jensen)
• predict quality of photos (Chinese or Italian) - Multilayer Perceptron
• association rules - AR; apriori
• forecast number of reviews - Prophet; ARIMA

Ethical consideration: Investment in restaurant using customer review indicators along with demographic, income, real estate data may be game changer for investors.

Project Document: Yelp_DataAnalysis

Google Play store survey Analysis

We are a hypothetical data science consulting firm and we have been approached by an app developer with a set of data regarding the India Google Play Store. They would like us to analyze the data and answer questions for them.

The development group is looking to create a hit app in India and has brought the following questions to the table:

Which apps are installed the most?
- **Which categories?**
- **Which age groups?**
- **Type (Paid vs. Free)**

Which are most popular in the last year?
Which are growing in popularity?
How can I make a "best selling"/highly installed app?
Which categories have the highest ratings?
- **Which are liked the most?**

How can I get more people to install an app?
What motivates people to write reviews?
Are ratings or reviews more correlated with installs?
Does size or installs have any effect on price?
Does price have any effect on size or installs?
Can we predict high total worth of an app with category, rating, reviews, size, pay type, and/or content rating?
Can we predict installs for a given category?

After reviewing the data provided we believe we can find an answer the bolded questions. We did not have the installation date so we could not answer "which are most popular by year" and "which apps are growing in popularity". We also didn't have review data (only review counts and ratings) so we were not able to determine what motivates people to write reviews or how to get more people to install an app. We did find some interesting anecdotal answers.

Technologies: R, ggplot, histogram, Hypothesis testing, Machine learning, Linear Regression

Techniques: data gathering, data mining, plotting, data transformation, descriptive analytics, predictive analytics, machine learning (kernel support vector machine), time-series, test statistics, data visualization

Ethical consideration: Real estate investment using market indicators or demographic data may worsen economic conditions for residents of that area.

Project Document: Google Play store Survey Analysis

MLB Tweets Sentiment Analysis

Social media has completely changed the way baseball is being marketed. Radio shows are supplemented with Facebook feeds and wall postings on the air. Every team, layer, reporter, website and organization has a Twitter feed updating constantly with news from the game and user interaction.

Stadiums used to run zany contests for fans; now they post trivia questions on Twitter for fans to tweet back an answer with their seat location to claim the prize. Rather than checking individual websites, fans are going straight to their Twitter and Facebook feeds to read news. Sometimes the news hit social media first, since there are less channels for the news to pass through.

Less is known, however, about the effect of tweets for team performance. Can tweets motivate or demote a team? Do fans have a sixth sense when it comes to team performance? To measure the effect of social media on team performance, tweets from a subset of Major League Baseball (MLB) games have been analyzed along with three hours of tweets tagged to teams prior to the game.

The tweet data were gathered, mined for sentiment, evaluated for “popularity” (by retweets and interaction with the candidate) and used in Multinomial Naïve Bayes and Support Vector Machines models to predict outcome.

Technologies: python, jupyter notebook, pyplot, seaborn, basemap, sklearn, Multinomial Naïve Bayes, Support Vector Machines, Grid Search Cross-Validation, Latent Dirichlet Allocation (LDA), Vectorization – CountVector, Ngram – unigram & bigram, Non-Negative Matrix Factorization (NMF), Vectorization – TF-IDF Ngram – unigram & bigram

Techniques: data gathering, text mining, plotting, data transformation, descriptive analytics, predictive analytics, machine learning, sentiment analysis, sentiment analysis on topics, data visualization e.t.c.

Ethical consideration: MLB can do various kind of data analysis on fans’ data/text collected from above mentioned applications and can plan for ticket discount for day time game or off season game. MLB can play jackpot based on users’ trend before game starts or many more. MLB social initiatives, such as to engage more fans or users through various apps, is appreciable. But it is also advisable to come up different offers to engage more fans towards increasing business.

Project Document: Jupyter Notebook,Twitter API and Sentiment and Topic Client

DATA Warehousing Project

After a few exercises in data science, I switched my focus a little onto database management and how data circulates in the process. Therefore, last semester, I took a data warehouse course. The course equipped us with how to build a data warehouse with ETL tooling in a variety of business scenarios. The project we did was to construct and implement two business processes using the fictitious Fudgemart and Fudgeflix databases.

Our team decided to answer following Business Processes and Value

1. Track (USA?) Regional Sales Fulfillment by Year per Department
2. Track Regional Fill Rate by Quarter by Sate
   FR: Monthly report by state indicating increases/decreases in plan types
3. Track Employee Hiring and Length of Employment Duration
   FR: Monthly report by department, supervisor, full-time status, hire/term date
4. Track Regional Distribution of Orders by Department
   FR: Daily report of products sold aggregated by department per USA region

In this data warehousing project, we initiated the project with enterprise bus metrics in order to build a star schema. In the bus metrics process, the sheet helped us to list the dimensions needed and the major fact tables. Next, we followed through the detailed dimensional modeling sheet to create the tables in the database using SQL scripts. Once the tables were established in the database, we designed an ETL source-to-target map to aid our actual ETL (Extract, Transform, and Load) process. With the map, we implemented the whole process in the Microsoft visual studio SSIS tool from source to stage, and eventually to data warehouse. In the end, we built an interactive business intelligence dashboard so that the business users could perform advanced analysis and business reporting.

Among all the projects I have participated in, this data warehouse project is probably the most time-consuming and complex because, unlike other data analysis projects, there were strict rules to follow. If we did not execute certain tasks as planned, it would take us more efforts and time to fix afterwards. Compared to data science, data warehousing requires lesser creativity but more consistency and attention to details. Generally, when it comes to data science, people instantly think of machine learning, modeling or algorithms. However, having at least basic familiarity of how data flows, and when and why some organizations choose to build a data warehouse environment is undoubtedly a great quality. It enhances my landscape towards data science, and boosts a better understanding of the data flow in order to help me make better decisions in the future. In my opinion, how the data is stored, transformd, and moved is tightly associated with data analysis because we need good data to produce accurate insights, but most people just choose to neglect. All in all, I would not say this is a typical data science related subject, yet it is just as crucial.

Project Document: Data warehousing documents

Conclusion and reflection

In summary, during my time with iSchool in Syracuse University, I had conducted many researches on various areas and topics, both individually and within a group. From my viewpoints, despite these projects being diverse from one another, the core principles of data science are similar and they can be broken down into below stages:

1. Define problem statement and goals;
2. Collect relevant data;
3. Exploratory data
4. Preprocess data
5. Engineer features and select features
6. Model development
7. Model evaluation and adjustment
8. Result interpretation and inference
9. Representation and
9. Business actions

This program not only cultivated my technical abilities, such as programming languages R, Python, statistics, and machine learning algorithms, but it also nurtured my soft skills, including teamwork, analytical thinking, and perceptive and open-minded. It also taught me to prepare a smart reliable research paper. I came into the applied data science master program with least knowledge about machine learning, yet I graduated from the program with abundant hands-on experience and expertise in many areas. From those experiences, I discover that knowledge is rigid, however, how you apply and interpret are powerful and agile. Though the professors taught me plentiful concepts, putting them into real-world applications yourself is far more beneficial. I learned to always remain curious and suspicious of what you know and what you see, and never jump to conclusions too quickly. It is vital to look at things with an open mind, and embrace what the data may lead you. Last but not least, when conducting a project, plan ahead but also leave space for imagination as the things do not always go as planned.