SYRACUSE UNIVERSITY

School of Information Studies

Applied Data Science Portfolio Essay

Laxman Kumar

LAXMAN KUMAR

315.880.8766 | lakumar@syr.edu
linkedin.com/in/kumarlaxman/ | github.com/Laxman-Kumar

DATA ANALYST

Data Visualization | Data Warehouse | SQL Database | Business Analysis| Machine Learning | Analytical Decision Making

EDUCATION AND CERTIFICATION

Master's (MS) Applied Data Science *(Minor Finance) (GPA 3.9/4.0)* | Syracuse University, iSchool | Syracuse, NY 2021

- Statistics, Data Warehouse, Big Data Analytics, Data Analysis and Decision Making, Information Visualization

Table of Contents

INTRODUCTION

"Data is a precious thing and will last longer than the systems themselves"
by Tim Berners-Lee.

The portfolio essay is written in an attempt to provide sufficient information that the objective set by the Syracuse University has been met successfully. It is a summary of my growth and experience in the last two years, which I gained at Syracuse University. Through the major academic projects, this portfolio will show my knowledge of data science and machine learning. During my master's, I took several courses from basic to advance level, which has fine-tuned my skills and prepare me to take on problems from wide domains - text mining, information visualization, data warehouse, database management, big data, and data analysis. Also, the curriculum has provided us the appropriate background, knowledge, training, and resources to become a successful data scientist. All the 7-objective set by the department have been met throughout the program. The explanation of each of the objectives will be buoyed with several final team projects completed throughout the curriculum.

One of the objectives focuses on collecting and organizing data. This objective was crucial for all the team projects completed throughout the curriculum. All the projects and assignments underwent honed my skills in data collection and data cleaning. Finding appropriate tasks and the relevant data was the most time-consuming initially. But as the semester goes, skill was mastered and now have enough knowledge to locate a relevant dataset from 100s of website. After locating a dataset, analyzing the dataset using statistical methods and supporting visualization is the primary task. During the last two years, I have gained expertise in various data visualization libraries such as matplotlib, seaborn, ggplot, and Plotly. All these libraries and knowledge of statistical analysis have helped me to master the third objective of the MS ADS program.

1. IST687 Introduction to Data Science

1.1 Project Description

The first project in which I participated during my masters was the Analysis of Airline Data. It was a group project. The data includes features like gender, age, flight information, airline class, flight route, satisfaction rating, etc. Customer satisfaction is identified through a likelihood to recommendation score filled in by the customer in their survey. This likelihood to recommend score is used to tag a customer as a Promoter, Detractor, or Passive. The Net Promoter Score (NPS) is then measured as the difference between % of Promoters and % Detractors. Detractors are the customers who are most likely to churn and may cause more customers to churn by spreading negative reviews. Currently, the airline has 14 partner airlines, of which 2 of the airlines have a negative NPS, which means that they have more Detractors than Promoters. The project's goal is to provide actionable insights and recommendations that will help increase their NPS. Moreover, South East Airlines needs help in deciding which partners to keep, which partners to drop, and which regional airlines should become their new partners. The Dataset received contains 10,282 customer surveys that capture several customer and flight attributes and the likelihood to recommend scores provided by each customer. In contrast, Promoters, on the other hand, may even bring in more customers by applying for positive reviews.

After cleaning the dataset, we analyzed the dataset using exploratory and explanatory data analysis. We created various charts using ggplot2 with explanation to completely understand the data. Then we used association rule mining to get the information on relevant attributes. After compiling the result of association rules mining, we moved forward to the modeling process. We applied Linear and Polynomial regression and performed hyperparameter tuning to increase the model performance. We also built a SVM multiclass classification model which is a nonlinear mapping for transforming data into higher dimension space. We documented the process and recorded the result into a pdf report.

1.2 Reflection and Learning Outcomes

The course gave me a solid understanding of performing data analysis and machine learning in R. It has allowed me to enhance my data analysis skills by using descriptive statistics and predictive modeling. Now I am confident of making a data analysis pipeline in R with all the necessary steps data cleaning, EDA, preprocessing and machine learning modeling. During this project, I gained knowledge about how model works, tested few hypotheses, cleaning data. preprocess data for modeling and finally train a model and interpret its output for business recommendations.

2. IST659 Data Administration and Database Management

2.1 Project Description

Knowledge of database management is critical to be a successful data scientist. Data Scientists are the people who make use of all the data stored in various forms of the database. We have trillions of petabytes of data stored in SQL and NoSQL format waiting for someone to analyze it and make use of it. Understanding database architecture and knowledge of extracting data from the database will always be crucial. Hence, to put together all the concepts learned in the class, we have created a *Missing Person Reporting Database Management System*.

Under the guidance of Prof. Hernando Hoyos, we could track down the small details needed to make a complete system. The motivation for choosing this domain for the project was because the world has a very high volume of child trafficking and missing people. Making a tracking system for the world with all the different law systems of countries was not feasible in the given time frame. So, we decided to go with India. India has a very high volume of child trafficking and missing people. Thousands of people go missing, and more than thousands of people are untraced and unidentified. With the lack of any centralized missing people database and poor infrastructure, it is tough to track and prevent. To track this problem, we have a database management system that can track a missing report made to the law officers. Our system can help complainers to get the flow of the missing report, from the complaint made at the police station to the steps taken to find the missing person. Also, the name of the officer assigned to the case is also shown in our system. We believe each person of the country should have an equal right to get their complaint taken seriously by law officers.

We used the following tools design the course of this project:

- Microsoft Visio: MS Visio was used to create the entity relation diagram for the project and conceptualize the attributes and relations of the different entities.
- Microsoft Access: MS Access was used to create the interface of the project system. We created forms which displayed necessary data to the user. We also used Access to generate reports
- Microsoft SQL Server: MS SQL server was used to create the database of the project. We wrote queries to make the various tables and inserted data using SQL. We also created views to using the same.
- Flutter: Flutter was used to make the mobile interface for the project system. Users can register themselves and generate complaints using their mobile devices or tablets.

2.2 Learning Outcomes

At the end of this project, I was confident in designing the database schema in MS Visio, building a database in MySQL, writing queries for CRUD operations and creating triggers and function to automate the process. Also, I learnt why relational database systems are important and are widely used. Also connecting MySQL database in MS Access and entering the data into the database using forms.

3. IST736 TEXT MINING

3.1 Project Description

I decided to take this subject for developing skills when dealing with unstructured and text data. This course focuses on preparing text data using tokenization, vectorization, lemmatization techniques. Textual data needs to be converted to proper vectors before feeding to any machine learning algorithm. As a course end project, I decided to analyze drug review data. The dataset contains approximately 53,000 reviews on more than 2500 different drugs. The dataset also includes two additional variables named useful count and date, which denotes the count of people who founds the helpful review and the date on which the review was submitted. The project's purpose was to attempt to predict a rating for drugs and sentimental analysis of reviews.

Our first problem will target predicting the rating of drugs based on their reviews. Generating sentiments about the drugs based on the reviews will also help people getting a 3- scale rating about the drug Positive, Negative and Neutral. Therefore, our second task is the sentimental analysis of reviews. Our third task is to predict conditions based on the reviews. This task is all about clustering, and we will only focus on the top 10 conditions in the dataset. Since the data is raw and fetched from online review sites, it must be cleaned before any analysis. Since we are also dealing with text data, the cleaning process may include cleaning the reviews using regex and removing HTML tags. One of the first and essential steps of data cleaning is handling null values. Also, as part of the data-cleaning process, we have renamed all the conditions with 'Not Listed' or 'Other' label from 'Others.' This will decrease the number of the unique group formed by the condition variable. Finally, all the reviews were filtered from a custom regex pattern to clean any non-alphanumeric characters such as emoticons or HTML tags.

We trained various classifiers (Linear SVM, Naïve Bayes and deep learning) and compared their accuracy to find the best model for a task. We have used accuracy, precision, and recall

comparing the models. I also incorporated several lexical and other non-word features such as review length in addition to adjusting vectorizer and classifier parameters. We have also removed rows based on drug and conditions whose reviews count is less than 2. Doing so will help us focus on the higher frequency drugs and conditions and, at the same time, increase our accuracy. After testing a different combination of vectorizers and its hyperparameter, the model's hyperparameter, we defined the best model for each of the tasks. We have also built a deep neural network for predicting ratings based on the reviews.

3.2 Learning Outcomes

From this subject, I refined my skills of mining data from social networking such as Twitter and preprocessing the raw data using regex and other data cleaning methods. The data fetched from social networking sites contain lots of data that is only used for texting and cannot be used to build a machine learning model. All the irrelevant data has to be removed before vectorization. I also learned how to tokenize data and use another method such as lemmatization to preprocess text-based data. In addition, I gained good hands-on experience with handling large documents and finding the most relevant topics using topic modeling.

4. IST718 BIG DATA ANALYTICS

4.1 Project Description

This is one of the subjects where we taught the concepts of Big Data using Map-Reduce and PySpark. Understanding of Big Data and Hadoop is necessary for a data scientist. As we have lots and lots of data coming every minute from tons of websites. Processing gigs of unstructured data, can we complete requires power of parallel processing using Hadoop or Spark. To test our knowledge and understanding using a group project, we decided to work hospital readmission data. It is a patient level dataset with a total of 101,766 rows representing 101,766 patients along with 53 columns representing 53 different patient attributes for each patient. An increase in Hospital readmission rates indicates poor hospital quality, inadequate treatment and results in the hospital getting penalized. Hence by identifying the factors that lead to higher readmission and being able to predict if a patient is going to be readmitted, the treatment provided by the hospital can be changed to avoid readmission, and thereby, the quality of healthcare provided to the patient can be vastly improved, as well as billions of dollars can be saved.

This project aims to help the hospitals accurately predict if a patient is going to get readmitted after discharge. The predictions made will help the hospital make informed decisions on the necessary treatment process alterations to reduce the patient readmission rate. By accurately predicting if a patient is going to be readmitted, the hospital will make necessary treatment changes to avoid patient readmission. This reduction in inpatient readmission rate will help the hospital save billions of dollars and improve its customer rating.

Through this project we got hands-on experience of building machine learning models by using pipelines, estimators and transformers available in the PySpark libraries. We got experience of working on datasets which are very big in size and contain a large number of rows and columns. This project also taught us the importance of using GitHub for version control and collaboration. The goal of this project was to accurately classify every patient under the 2 classes 'Will be Readmitted' and 'Will not be Readmitted' by building predictive machine learning/deep learning models which are capable of doing binary classification. Through our methods, processes and modelling techniques we were able to accomplish this goal with a 64% accuracy in our prediction, i.e., out of 100 patients, 64 of our patient predictions are most likely going to be correct.

4.2 Reflection and Learning Outcome

From the course, I learned the fundamentals of Hadoop and map-reduce. This course was taught in PySpark, and hence I gained hands-on experience with PySpark also. Experience with PySpark is necessary to become a data analyst. Sometimes we might deal with massive data, and with the parallel processing power of PySpark, dealing with the data will be easy. I also studied to build a machine learning pipeline in PySpark. Overall, this course has helped to make a step forward towards the data science career.

5. CIS700 MACHINE LEARNING FOR IOT DEVICES

5.1 Project Description

In this course, we studied many research papers and some high-level concepts when dealing with sensor data using deep learning. I also learned to read a paper and locate the strong and weak points in the paper. In the end, I joined the professor for one of the researches. The research focuses on speed detection using accelerometer data. Speed detection is not an easy task, and it contains multiple challenges. First, there are limited data, and not many organized data are generated by an accelerometer in different walking or running speed. Second, different walking speeds have salient features that a machine learning model needs to extract.

With the above challenges, it is hard to get a large amount of perfect and clean data for the machine learning model to train. We ask our participants to walk on a treadmill to control the speed and collect the accelerometer data for different rates separately to collect speed data. However, the accuracy of neural networks won't be good without a significant amount of data. We will introduce a prototypical network, a few shot learning approaches used to train the model.

Recently, lots of methods for few-shot learning have been discovered. One of them which has gained high popularity and has an outstanding result is meta-learning. The idea of prototypical learning has come from meta-learning, which is similar to k-means clustering. Prototypical learning learns to differentiate a small number of examples with fewer data provided. Similar idea to traditional CNN, prototypical networks divides the data into a training set and testing set. However, the prototypical network will take N number of classes and take K sample known as N-way K-shot, where K is typically small. More will be introduced in the experiment section. We create a CNN with 14 classifiers to classify the steps as the baseline; the accuracy is low (around 45%). Simultaneously, the prototypical network will be a way to solve the problem in which the accuracy can reach about 85%, which is the novel part of the experiment.

Besides speed detection, we've also done experiments on classifying walking steps between driving data and other data which we ask our participants to wear the accelerometer during their daily life, which we found it is easy to mix walking data with driving data to identify whether a step is walking or non-walking, we implement a CNN model for the experiment. Since we have both waist and wrist data, we will experiment with three different models and two different classifiers. The first classifier is a binary classifier, which divides the steps into two groups, walking, and non-walking. The second classifier is a three-class classifier, which is walking, driving, and others. We also build three different models for two classifiers: models that consider only the waist or wrist data and a combined model that concatenates the waist and wrist CNN model with a linear layer. Overall, we tried more than 50 different models and hyperparameter combinations to get the best result possible.

5.2 Reflection and Learning Outcomes

From this course, I studied how to read paper efficiently and track weak and strong points in the paper. I learned about some of the new sensors that recently hit the market and can get the human reading. Also, we scratch the surface of some advanced deep learning models such as few shot learning, encoders and decoders, CNN, image processing and RNN. I participated in a research and got research experience as well from this course.

CONCLUSION

The portfolio essay reflects my work and learning during my journey of pursuing a master's degree at Syracuse University. I believe my journey has been productive as I learned a lot in the last two years. I had successfully built a strong foundation which will help me start my career as data scientist. Also, I enjoyed a lot during last years I made great friends and participated in different activities.

GITHUB LINK

https://github.com/Laxman-Kumar/MS-ADS-PORTFOLIO.git

VIDEO URL

https://youtu.be/UVGofCNSurs