**Portfolio document**

**Minyang Wang**

The study in the Applied Data Science program in iSchool was one of the most important and precious experience in my life. During this one and half year program, I have not only possessed the knowledge and skills of utilizing data science techniques to solve real-world problems, but also succeed in determining my career goals. Before I entered the ADS program, I first gained my bachelor’s degree in economics in Renmin University of China and then I came to Syracuse University to study MPA program in Maxwell school. It was in that spring semester I took the IST 687 as my elective course, and I gained strong interest in the data science field. I found it was so amazing to use the data and algorithm to solve the problems in the real life. I also realized that in the future, no matter in which industry, in private sector or public sector, the demand of data analysis will grow dramatically. Then I applied the ADS program and this one and half year of experience tells me that my decision was correct.

In my portfolio document, I will introduce 6 projects that I have done during the program. I will talk about what I have learnt from the projects and how the projects demonstrate my mastery of the learning objectives for the program.

**Project 1: Hyatt Hotel Analysis: Recommendations to Improve Net Promoter Scores**

**Course: IST687 Introduction to data science**

**Course description:** This course provides applied examples of data collection, processing, transformation, management, and analysis as well as a hands-on introduction to the emerging field of data science.

**Project description:** The purpose of this project is to demonstrate and exhibit a high-level analysis based on the data gathered from Hyatt and their customers ratings. Each group was asked to analyze the Hyatt customer data and provide suggestions to increase the NPS (Net Promoter Score) score of the hotel based on their analysis.

**My gain from this course and project:** I took this course when I was still pursuing my first master degree in Maxwell school. Because of this course I decided to continue my study in the data science area. In this course I first got into touch with data analysis and started to deal with data by using R. I learned to use R to preprocess the data such as removing NAs, to do the descriptive analysis such as aggregating dataframe and plotting graphs, to use different algorithms to do the machine learning analysis. In the course project I got the chance to incorporate the theoretical knowledge with the real-world problem which makes this course more practical. It was my first time to deal with such large scale of data with more than 10 csv files of which each is more than 1 GB. I learned to clean the dataset and reduce the number of columns according to our specific business questions. I also practiced the algorithms I learned from the class to apply them on the data. Thanks to this project, I got to know what a data analysis project is like and what a data analyst does for their job, and I determined to become one of them.

**Project 2: H1B Application Result Determinants Analysis**

**Course: IST707 Data Analytics**

**Course description:** This course will introduce popular data analytics methods for extracting knowledge from data. The principles and theories of data analytics methods will be discussed and will be related to the issues in applying data analytics to problems. Students will also acquire hands-on experience using state-of-the-art software to develop data analytics solutions to scientific and business problems. The focus of this course is in understanding of data and how to formulate data analytics tasks in order to solve problems using the data.

**Project description:** the goal of this project is to analyze the determinants of an approved or denied Labor Condition Application case and to predict the case status of an application submitted by the employer to hire non- immigrant workers under the H-1B visa program, so that based on the factors which may affect the result of an application, the employee can choose the employers who have the highest possibility of receiving approved LCA.

**My gain from this course and project:** This course is the advanced level of IST687 where I learned the theory of each machine learning algorithm. From the IST687 I learned what each algorithm is and how to apply them to the analysis, but from the IST707 I learned why we use this algorithm instead of that one and how to tune the model to get better analysis results. For my course project, I faced a most common and troublesome problem: my data was highly biased that more than 95% of my data was categorized to one label (95% of the application were approved whereas there were only 5% denied cases). This made it very hard to build the prediction model and to evaluate the model because the baseline is 95%. So from this project I got better understanding of each machine learning algorithm and learned to deal with highly biased data.

**Project 3: Curriculum Information Sharing System Design**

**Course: IST659 Database Administration Concepts and Database Management**

**Course description:** IST 659 is an introductory course to database management systems. This course examines data structures, file organizations, concepts and principles of database management systems (DBMS); as well as, data analysis, database design, data modeling, database management and database implementation. More specifically, it introduces hierarchical, network and relational data models; entity-relationship modeling; basics of Structured Query Language (SQL); data normalization; and database design. Using Microsoft’s Access and SQL Server DBMSs as implementation vehicles, this course provides hands-on experience in database design and Implementation through assignments, lab exercises and course projects.

**Project description:** This project is supposed to build a course information sharing platform by using Microsoft SQL for Syracuse University graduate students to help students better select courses before the start of each semester.

**My gain from this course and project:** In this course I learned the knowledge of database management from designing the Entity Relation Diagram based on 1NF, 2NF and 3NF to using the SQL to build database and to query the data. The course project is the combination of the content of this course through the semester. I learned to design the entities of the database and to build relationship to connect each entity. Then I wrote SQL codes to build my database and to query the data to answer questions which are related to the purpose of my project. To present and visualize the database I also learned to use Microsoft Access to create the user interface to make my project more like a business product.

**Project 4: My Alipay Annual Bill Poster**

**Course: IST 719 Information Visualization**

Course description: This course introduces students to skills and techniques related to information visualization. In this skills-based course, students are introduced to the R programming language, Adobe Illustrator, simple data cleaning techniques, simple design concepts, different visualization tools and the ethics of visualizing data. The focus is on developing static data visualizations to visually explore and communicate findings using data from a variety of sources. Conceptual themes are presented alongside technical aspects of data visualization.

**Project description:** Alipay is a third-party mobile and online payment platform established in Hangzhou, China in February 2004 by Alibaba Group. It overtook PayPal as the world's largest mobile payment platform in 2013. Users can use it not only for online payment and account transfer like PayPal, but also for offline shopping. In this project, I downloaded my Alipay transaction history data since I created my account to analyze my consumption behavior in the past few years and visualize the information by R and Adobe Illustrator.

**My gain from this course and project:** The data visualization skill is one of the most important skills that a data expertise should have because we need to present the analysis results to the audience who generally don’t understand the term and jargons of data science. From this course I learned to use R to all kinds of plots and to use Adobe illustrators to revise and adjust the plots.

**Project 5: Gstore Revenue Prediction**

**Course: IST 718**

**Course description: Big Data Analytics**

This course is a broad introduction to modern techniques in data science including elastic net regularized regression, random forest, gradient boosting, and deep learning. It emphasizes a statistical learning point of view, and a careful examination of generalization error, model interpretability, feature engineering, and bias-variance tradeoff.

**Project description:** Learning the customers’ behaviors and their abilities to purchase is always one of the most important methods for the merchant to increase their profit. This project stands at the point of Google Merchandise Store (Gstore) trying to maximize the profit by predicting the revenue that each individual user may contribute to Gstore.

**My gain from this course and project:** This course reaches to a much higher-level data analytics knowledge and I learned more mathematical principles of each machine learning algorithm and had a better understanding of the similarities and differences among algorithms. Instead of coding on R, this course required me to do all the analysis on Python that practiced my programming skills in a different language. The project is a process of finding problem and solving problem, from unable to upload large cvs file to Python, to Python shutdown due to the massive computation when running for loop to tune the model. By searching the solutions to these problems online, I learned much more than what the professor had taught in the lecture. I did acquire the ability of problem self-diagnosis and problem self-solving, which are the most important basic skills that a data analyst should possess.

**Project 6: Japanese Whisky Review Analysis**

**Course: IST736 Text Mining**

**Course description:** Introduces concepts and methods for knowledge discovery from large amount of text data, and the application of text mining techniques for business intelligence, digital humanities, and social behavior analysis

**Project description:** Over the last century, globalization allowed economic trades to circulate between countries. Native country products can be sold to the other side of the world. Japanese whisky as one of the most representative Japanese products is beloved by the western world. Among all type of whiskeys, whisky sellers need to know which type are more favored by the customers. Instead of reading through every words of comment, a more efficient analysis method is adopted – machine learning text mining algorithms. To extract the key words from the huge amount of text, topic modeling and classification mythologies are applied for analyzing the whiskey reviews. By applying text mining knowledge, the business question “How can Japanese whiskey companies increase their whiskey online sales?” can be converted to “What are the feathers of different whiskey brands reviews?” and “What are the feature words of positive and negative reviews of different whiskey brands and bottles?”

**My gain from this course and project:** After taking this course, l learned the basic concepts and methods in text mining, such as text representation, text classification clustering, and topic modeling. I also learned the similarity and difference between text mining and data mining. The project incorporates the text mining concepts and methods to model real-world problems and by doing the project I got familiar with the whole process of text mining analysis. I learned to use python to do the corpus descriptive analysis and learned to use machine learning algorithms (Naïve Bayes, Support vector machine) to train the model to predict the sentiment of the whiskey review. Considering the growing demand of NLP expertise, the acquisition of text mining knowledge can make me more competitive in the job market.

The ADS program expects us to be able to describe a broad overview of the major practice areas of data science; to collect and organize data; to identify patterns in data via visualization, statistical analysis, and data mining; to develop a plan of action to implement the business decisions derived from the analyses; to demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization. All the projects from the courses of the program are designed to help students to practice such skills.

The Gstore revenue prediction project of course IST 718 can best demonstrate how I practice these skills. At first I got the data and came up with business questions which can be solved by the data science concepts I have learnt from the class. We decided to predict the potential revenue of a user visiting the Gstore website. Then I cleaned the data according to our business questions by selecting the columns that are needed for the analysis, and removing the irrelevant columns and rows which has repeated data or NAs. When doing the analysis I used the algorithms learnt from class such as random forest algorithm and gradient boosting tree algorithm, and tune the models to get the best performance. Finally we visualize the outcome of the analysis by making a poster and present the poster to not only professor but also the guests visiting Ischool on the poster day which practiced my data visualization skill and communication skills regarding data and its analysis for different type of audiences.

I think the different kinds of projects are one of the best features of the ADS program. By doing the projects I learnt the practical and useful skills that will help me achieve success in my career life. I will always remember the days in the ADS program and I will always take pride of being a student in iSchool.