Portfolio Milestone

Author: Wanyue Xiao

SUID: 720633297

Academic Email: xwanyue@syr.edu

LinkedIn Website: www.linkedin.com/in/wanyue-xiao-60688196

Table of Contents

Introduction

Author Information

- Created by: Wanyue Xiao
- SUID: 720633297
- Academic Email: xwanyue@syr.edu
- Github Repository: https://github.com/xwanyue0221/Graduate-Portfolio
- LinkedIn Website: www.linkedin.com/in/wanyue-xiao-60688196

Current Resume

The latest version of professional resume could be procured through the available link, which was updated in March 2021.

Link: https://github.com/xwanyue0221/Graduate-Portfolio/blob/main/wanyuex.pdf

Self-Introduction

As a master program held by the School of Information Studies at Syracuse University, Applied Data Science (ADS) instructs students with skills of extracting, analyzing, visualizing, managing, and storing data to create insights using various tools and techniques. Correspondingly, courses taught by experts in different domains are available for students to attend, ranging from data analytics to database management. Some of those courses emphasize the value of teamwork and the importance of collaboration while some of them value individual competency in terms of programming, academic project reporting, and oral presentation. According to the program's instructions, a competent and qualified graduate student of the ADS program should not only have the ability of data processing and data management but also be able to gain business insights from data analysis and to demonstrate communication skills with stakeholders. During the 2 years of studies, I took courses in data acquisition, database management, data visualization and analytics, natural language processing, business decision making, and financial management. In each of those courses, I finished required assignments, quizzes, exams, and team projects, through which I master different areas of expertise and skillsets (such as R, Python, SQL, NoSQL, and Tableau). Therefore, this paper is a graduate portfolio milestone for my 2-year-master program, including the details of 6 major academic projects from the courses listed above and a self-reflection on them.

Academic Projects

Semester	Course	Project Description	Tools & Skills
2019 Fall	IST 659 Database Administration	**ToyPedia Online Forum Database Design** This project is to re-design a database system for an online non-profit group called ToyPedia, whose tasks include user requirement analysis, ERD model design, database implementation, sample data testing, query testing, temporary user interface creations.	**Language:** SQL **Tools or Skills:** Microsoft SQL Server, Data Modeling & Design
2020 Spring	IST 719 Information Visualization	**Video Games Sales Analysis Infographic** This project is to create a poster for a real-life problem or an authentic dataset, utilizing data processing techniques and information management knowledge obtained from textbook. The chosen topic is video games sales dataset.	**Language:** R **Tools or Skills:** R Plotting Packages, Adobe Illustrator, Poster Creation
2020 Spring	IST 707 Data Analytics	**US Car Accidents Severity Analysis** This project is to unearth causes of U.S.A. nationwide vehicles accidents happened between 2016 and 2019 through statistical analysis and machine learning analysis, seeking potential association relationship or causal relationship between accident severity level and 48 exterior elements (such as weather and road conditions)	**Language:** Python, R, **Tools or Skills:** EDA, Statistical Analysis, In-depth Machine Learning, Scikit-Learn
2020 Fall	IST 718 Big Data Analytics	**Telecom Customer Churn Analysis** This project aims to discover critical features influencing churn rate, and then to develop models that are capable of predicting customer churn after comprehensive data exploration and analysis. All the implementation and codes were executed in PySpark environment.	**Language:** Python **Tools or Skills:** Apache Spark, Machine Learning, Statistical Learning, Time Series Data Analysis
2020 Spring	IST 664 Natural Language Processing	**Detection of Personal Experience or Storytelling** This project is to find the appropriate classification method to predict the existence of personal experience or storytelling and to summarize potential linguistic features that are crucial for the detection results based on the quantitative analysis of Tagged Reddit Submission Dataset.	**Language:** Python **Tools or Skills:** Statistical Parsing, Semantic Processing, Role Labeling, Discourse Analysis, Coreference Resolution, Deep NLP
2020 Fall	IST 700 Deep Learning, NLP, and Computational Social Science	**COVID-19 Sentiment Analysis Report** To gain more valuable understanding of public opinions and support drafting actionable policies, this report aims to find topics of English- language COVID-19 related tweets posted in April and further to explore variations in how the COVID-19 related tweets, topics, and associated sentiments changed over a period of time.	**Language:** Python **Tools or Skills:** Data mining methods, Topic Modeling, Corpus Annotation, Word Embedding, Transfer Learning

Database Management

I took two courses specifically for stringing my ability in data management, including IST 659 Data Admin and Database Management and IST 776 Advanced Database Management. The former one is an introductory level course of database management systems which mainly focus on relational database design and implementation. As the advanced class of IST 659, IST 776 foreground the implementation of non-relational database management systems such as Hadoop Ecosystems, MongoDB, Redis, and Cassandra.

IST 659: Data Admin and Database Management

After finishing IST 659, students should be able to comprehend data concepts, select appropriate database development cycles if applicable, use data modeling and data normalization skills during database construction, create database and database object using popular SQL management tool, evaluate the efficiency and effectiveness of databases thoroughly, and adjust database's strategies to satiate customer's requirements.

To utilize and testify the knowledge gained from the course and labs, students need to discover a real-world database problem and then provide a feasible database solution targeting this question. The final project is a 2-person group project where I and my groupmate designed a comprehensive database solution for an online web-based forum called ToyPedia - it is a small forum developed by Yan Lee in 2012. The original forum only had fundamental functions such as posting and replying and user management. We conducted a thorough investigation of user requirements and found problems hindering the effectiveness of the original database – those problems include issues of an increasing number of users, database incompatibility, data redundancy, limited storage capacity, and poor maintenance. Therefore, we proposed a new database system and drawn out a an ERD model (Fig 1.) and a data flow according to its business process. To satiate the needs, we added several new functionalities into the new relational database – it includes revised posting and replying functionality, user management with access management merged in, product selling and purchasing management, and payment and delivery management. Using this database, we expected to improve effectiveness in different aspects. One example is that the database should perform query function and return responds quickly when relevant information needs to be retrieved. Another detailed example is that simultaneous update of data on multiple devices is feasible.

From a technical aspect, we used the Microsoft Visio software to draw the ERD model and used the SQL Server Management Studio to create physical tables according to the logical plot. Reports and stored procedures were created to display a screenshot of all the user information, the purchase records of each user, items listed in user's shopping cart, the inventory level of each product, the best-selling products in a certain period, and revenue gained during certain period (Fig. 2). All of these questions proffer business insights to the management team. To further reduce the redundant workload that might happened in the process of user information management, a trigger had been created to automatically assign coupon to new registered user. Finally, a prototype with user interfaces was completed to prove that the database is functionable.

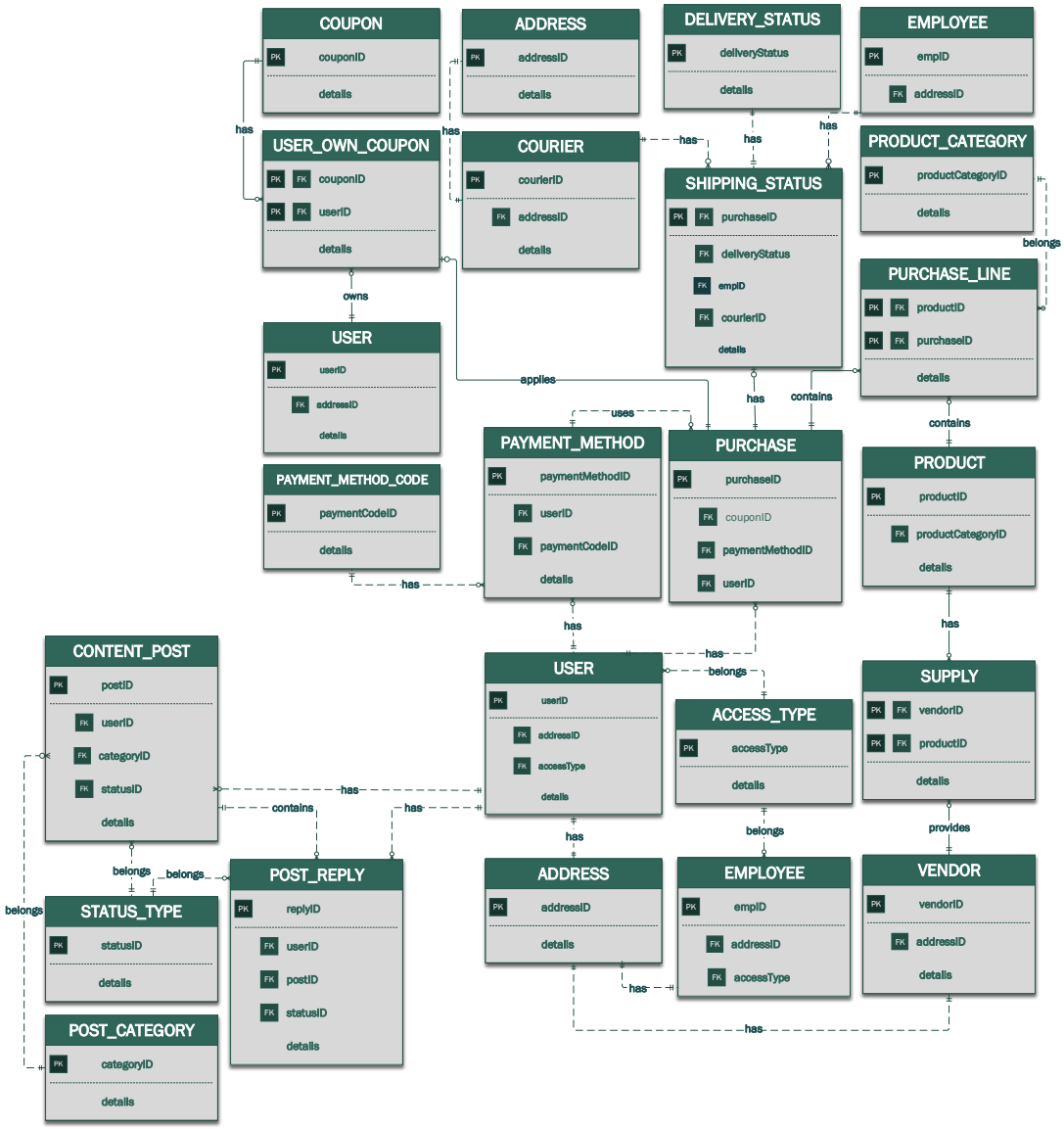

Fig. 1: Logical Model, IST 659

User Expenditure Report

UserID	UserName	NickName	User Type	Expenditure
U0001	Rohit Menon	Appu786	User	$3,149.42
U0002	Lawton Ray Zavier	ZavierL	VIP User	$5,143.35
U0003	Sonamgyalpo Sherpa	Sonam	User	$2,348.30
U0004	Prateek Prabhakar Reddy	Pnprabha	User	$4,141.20
U0005	Rachit Kaushik	MaxImum	User	$2,587.08
U0007	Shaw Wan	Poppy	VIP User	$1,274.52
U0008	Rousseau Michael	RM001	User	$2,668.38
U0009	Shinnick Ming	M&M	User	$692.28
U0011	Anakin Skywalker	Skywalker_A	User	$2,068.60
U0012	Thomas Shelby	Thomsa	User	$288.24

Fig. 2: Stored Process of User Expenditure, IST 659

Data Analytics and Visualization

To practice skills in data analytics and data visualization, I took course IST 719 Data Visualization, IST 707 Data Analytics and IST 718 Big Data Analytics. The former course taught me all sorts of techniques required for visualization, including cleaning and exploring data, data aggregation, simple design and information organization skills, and quality graphic presentation of data visualizations. From the last two courses, I learned prevailing data mining methods, comprehended theories and implementation of statistical models and machine learning models, acquired hands-on experience using edge-cutting software. I applied all the models and skills listed above on various datasets and evaluated these models in different perspectives. Therefore, all knowledge gained from these two courses enable me to deal with problems with great flexibility.

IST 719: Information Visualization

The deliverable of IST 719 is a poster project that leveraged skills developed throughout the semester. In this individual project, I collected data on the records of video games with cumulative

sales greater than 100,000 copies from 1980 to 2015. The dataset contained several variables, including the information of each video game, the publisher and platform, the year of publish, the genre, the ranking, and sales number in the region of North America, European Countries, Japan, and other countries respectively. First of all, the target audiences I identified for this dataset were game developers who lubricate from the video game market and people who have a passion for and fascination with video games. The identification of the target audience helped me to narrow down the scope of the investigation. Next, I created five plots that reveal the hidden relationships or patterns among those variables using R and Adobe illustrator. Two of them are single-dimensional plots that demonstrate the top 10 publishers while the rest three are multi-dimensional plots that show the sales distribution by different game genres, the sales distribution in these four regions, and the distribution of game genre in 30 platforms (Fig. 3).

What I learned from this project is that knowing which ones to use and when is essential. For example, to visualize both the distribution of game category and the sales records by each genre in different regions, I cherry-picked a heat map that could show differences in data through variations in color and size. To improve the design aesthetic, I selected the Voronoi-Tree map – a variation type of heat map – and modeled it into a circle (Fig. 4). What's also enlightened me is that using a different combination of colors, the plot is effective in showing changes and trends and conveying graphic information that is interpretable for the layman. On some occasions, even though those plots failed to present explicit ideas, they provide a hint or a direction for future investigation. Challenges are inevitable when one is trying to solve a real-world problem, ranging from data preprocessing to poster design. In the progress of data preprocessing, I needed to decide the approaches of error management – replacement or removal. Apart from the errors, different plots required disparate data types, indicating that the step of type conversion was unavoidable. What's more, to find out the optimal combination of colors used in posters, I browsed the Internet for hours and looked at other experts' dashboards or posters for inspiration.

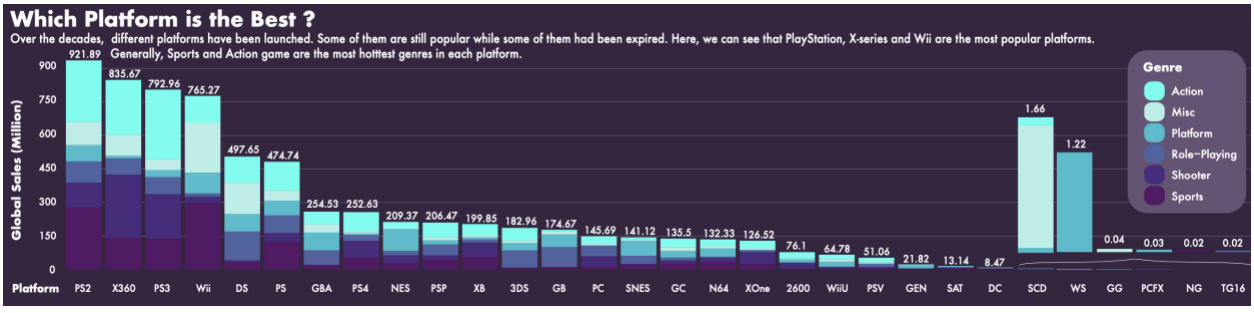

Fig. 3: Top 6 Game Genres in Different Platform, IST 719

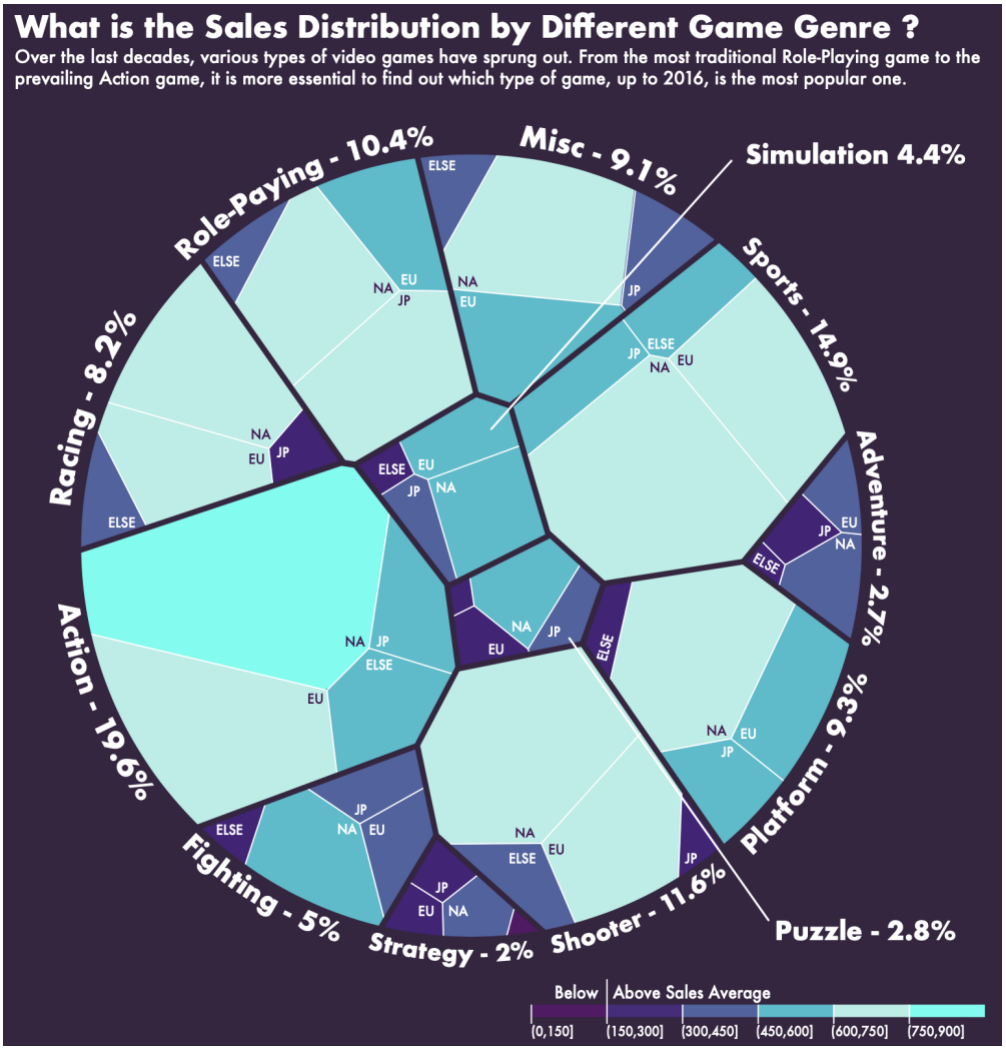

Fig. 4: Sales Distribution by Different Game Genre, IST 719

IST 707: Data Analytics

The final project of IST 707 is a 3-person group project. The deliverables included a project idea proposal, code files that implemented, trained, and tested text mining models, and an academic project paper that detailed the data mining problem, the significance and impact of such problem, the implementation and evaluation of text mining approaches, and the interpretation of discovered patterns. In this project, our group wanted to have a better understanding of the causes of car accidents. My teammates and I randomly sampled 50 thousand records with 49 variables from 3.0 million car accidents that happened in 49 states in the United States, starting from February 2016 to December 2019. Through the analysis, we anticipated finding the latent association between the

severity level and the rest, which include the location, time, weather, traffic, and report of the accidents. In this research, our team applied seven machine learning algorithms (Association Rule Mining, Logistic Regression, Decision Tree, Support Vector Machine (SVM), K Nearest Neighbor (KNN), Random Forests (RF), XGBoost) to extract the most significant features affecting the severity of car accidents and conducted an unbiased and low-variance cross-validation for each algorithm. The project began with data munging, in which we replaced the missing values using the KNN Imputation algorithm and removed outliers. In the process of feature engineering and feature selection, data-type conversion, one-hot encoding, and standardization were implemented. The following process is model training and validation. To procure some empirical results, we trained an Association Rule Mining model, from which we realized that these accidents most likely happened in the daytime. For the rest models, six different models with careful hyperparameter tunning were applied to the dataset. Finally, based on the feature coefficients and weights generated by each model, we could generate insights about the research question. This project enabled me to familiarize myself with the data science process, helped me to compare models from different aspects, and instructed me to interpret models' results based on the proposed question.

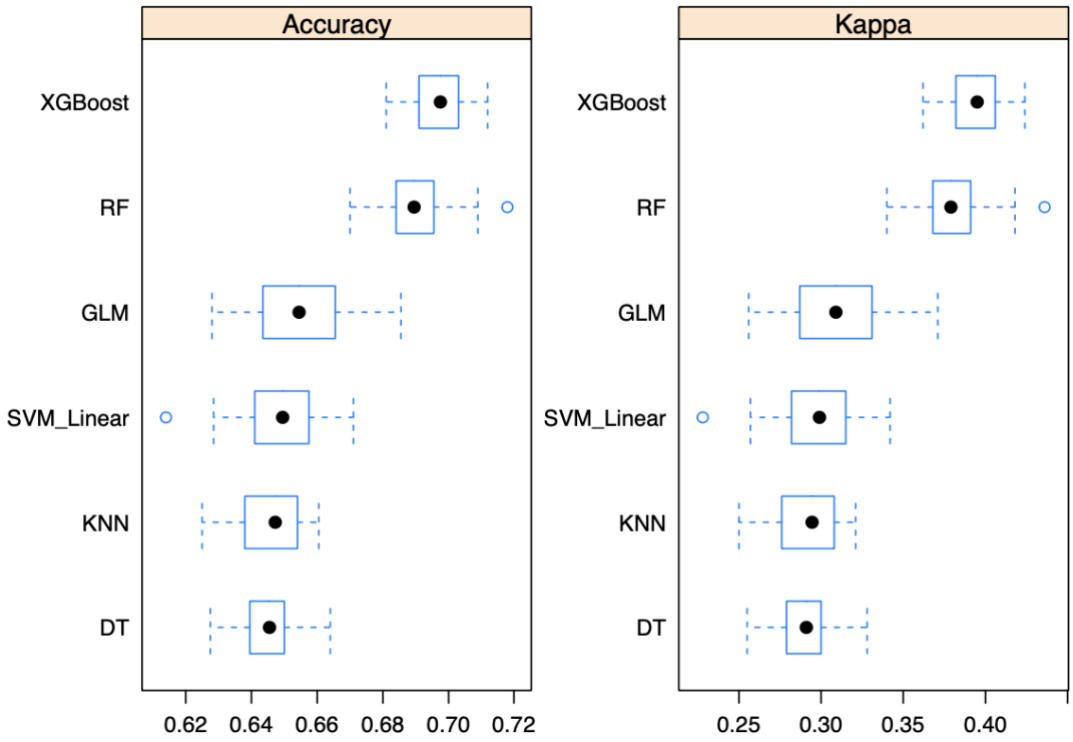

Fig. 5: Accuracy Comparison of Six Classifiers with 10-fold Cross-validation, IST 707

Among these models, Gradient Boosting Machine (GBM) has the best performance (Fig. 5). According to the comparison results, road condition, weather, and TMC code are the three most

critical features for classifying accident severity. No traffic signal, no crossing, and no station in the nearby location tend to cause a severe car accident. A lower pressure between 29.5 and 30.2 inches, which denotes stormy weather, is more likely to be associated with high severity level accidents. Therefore, we concluded that the severity level of traffic accidents could be predicted by exterior conditions like road conditions and weather. Allowing the readers to adjust all of these models, we created a Shiny App that could be secured through the available link (https://yiyuan-cheng.shinyapps.io/IST707_Final/).

IST 718: Big Data Analytics

IST 718 is an advanced data analytics course, assuming students have already equipped with quantitative skills and strong programming skills. As for the final project of IST 718, it is a 3-person team project, for which the whole project was run on Apache Spark environment and databricks. Our team wanted to investigate the relationship between customer churn rate and other customer's information in the Telecommunication industry. My teammates and I collected a telecom customer churn dataset consisting of 7043 rows and 21 columns. Along the data science process practiced in IST 707, our team proposed several inferences, wondering how the number of services registered influence customer churn, how personal characteristics affect customer churn, and how changes affect customer churn. Therefore, we predicted customer leaving based on the customer's information, the account's information, and the services they signed up for by building machine learning models (K-Means Clustering, Logistic Regression, Linear Support Vector Machine, Random Forest Model, and Gradient Boosting Model).

Before starting the process of modeling, we used several different preprocessing methods such as index variable dropping, data type transformation, missing values replacement, principal component analysis, and class imbalance management. Besides, due to a large number of features, the technique of dimensionality reduction or feature selection, in which case the results of exploratory data analysis would be informative. One of the plots we created is a correlation map (Fig. 6) which could be beneficial for key feature identification. Accordingly, a strong correlation (which has absolute values above 0.3) could be found between the target variable and Contract, OnlineSecurity, and TechSupport respectively. It is worthwhile to point out that TotalCharges and MonthlyCharges seem to have a rather strong correlation, which eventually will cause multicollinearity issues. Therefore, Variable Inflation Factors (VIF) could be used to detect the

strength of the correlation of a variable with a group of others. What's more, churn correlated with tenure and payment through electronic checks.

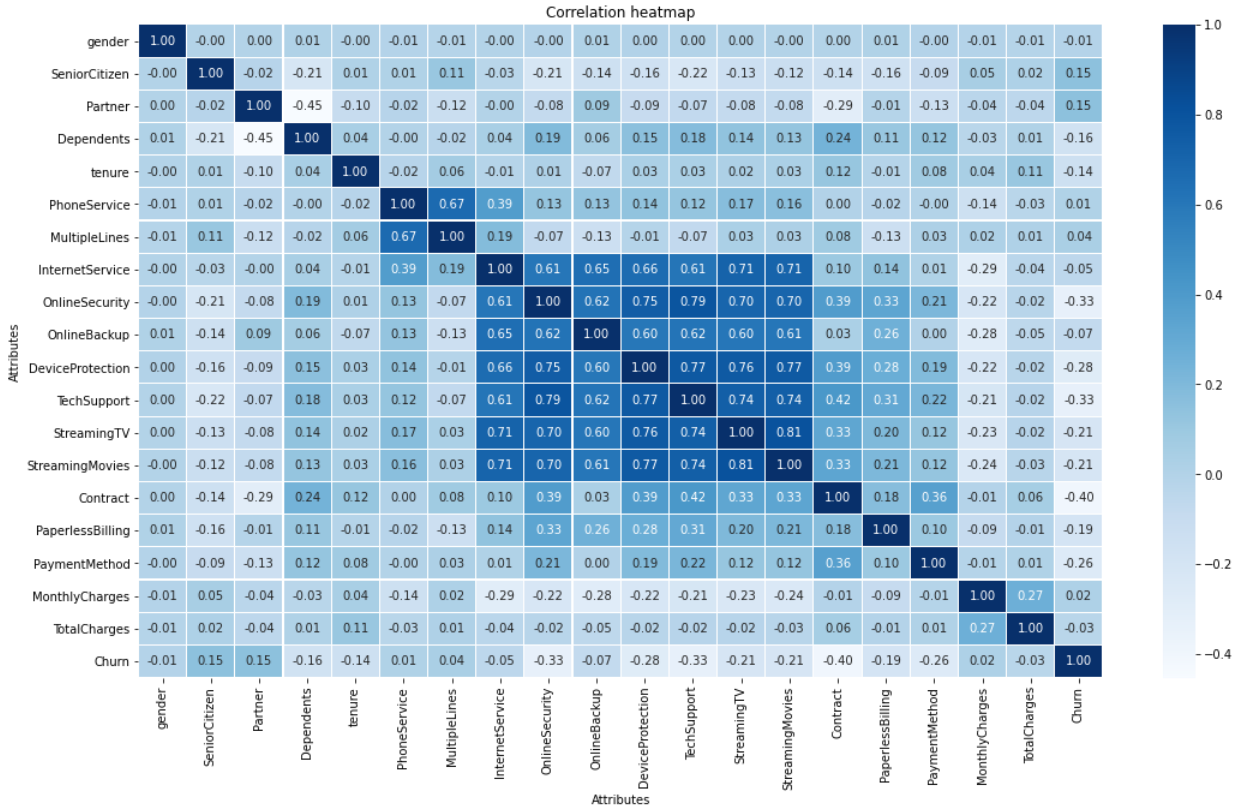

Fig. 6: Correlation Map of Customer Churn Dataset, IST 718

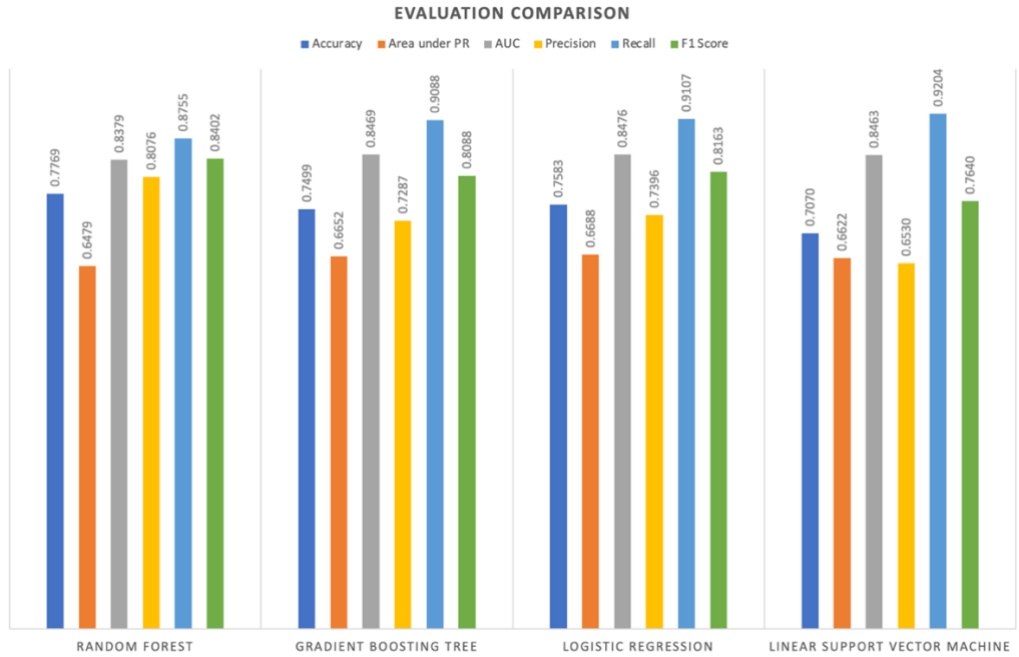

Fig. 7: Evaluation and Comparison of Six Classifiers with 10-fold Cross-validation, IST 718

This project successfully developed, concluded, and evaluated these four prediction models mentioned above (Fig. 7), providing comprehensive suggestions to the proposed questions. The result indicates that customers with Online Security and Tech Support Services prone not to default while those customers with characteristics of a partnership and longer contract tend to stick with the company. Even though this study used coefficients as an indication of feature importance, the result still has the possibility to varied until a significance test and t-test has been conducted.

The takeaway note of this project is that simply pursuing a higher accuracy rate during modeling and hyperparameter tunning would be undesirable given the potential issue of overfitting and lopsided interpretation. Another experience I gained from this project is that finding hidden business patterns and generating business insights that are beneficial for decision-making are as important as building models with the best performance. For example, the reliable indicators of a loyal customer would be a longer tenure and a registration on the online security or tech support service, whereas a churned customer is more indicated by having a monthly contract. Services such as tech support, online backup, and online security are inferred to be an advantage to the company. Therefore, the telecom company knows that a length contract, a recommendation in registering featured services, and an automatic payment would be crucial indicators for recognizing loyal customers.

Natural Language Processing

Text mining is an essential part of artificial intelligence technology, utilizing natural language processing to transform unstructured text into normalized and structured documents. Text analytics was a brand-new area that I never stepped in. With interests in sentiment analysis, I took IST 664 Natural Language Processing, IST 700 Deep Learning, NLP, and Computational Social Science, and IST 736 text mining. IST 664 is designed to help students developing techniques of unstructured text processing, including concepts regarding text analysis and the ability to generate a linguistic analytic report, whereas IST 736 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. IST 700 is an advanced course of IST 664, aiming to introduce advanced, prevalent data mining methods for extracting knowledge from unstructured data.

IST 664: Natural Language Processing

The final project of IST 664 was challenging. Comparing with these projects I have participated in before, this project focused on solving problems from an academic perspective instead of from a business perspective. The goal of this project was to detect personal experience or storytelling based on a dataset annotated by Professor Lu Xiao and her research team. The dataset is from a research paper named 'The Art of Justifying in Social Media: Insights from Reddit "Change My View" Submissions'. There are 330 submitted comments retrieved from Reddit Discussion Posts and corresponding personal experience marked manually to indicate whether there is a personal experience, valued 1 if there is and 0 otherwise. One pattern we observed from the dataset is that the sentence containing personal experience was mostly conveyed in a few successive sentences. After removing all the anomalous data, merely 8,468 sentences have remained. Based on the dataset available, deep learning or neural networks are out of our focus since their performances would be passable if the dataset was not scalable. Instead, we performed the statistical analysis and trained an SVM model on the same dataset.

Firstly, we performed word frequency analysis to get the general image of the dataset by analyzing the frequency of the top 100 words, yet we found that most words were uninformative and overlapping. Then, we applied the Naive Bayes algorithm to predict words with personal experience, achieving 0.5859 of accuracy rate. Additionally, we performed a statistical analysis by comparing the manual annotated values with automatic marked values. In this step, the Stanford CoreNLP library was utilized to detect past tense sentences while regular expression was used to match the sentences with VBD and PRP words with the regular expression.

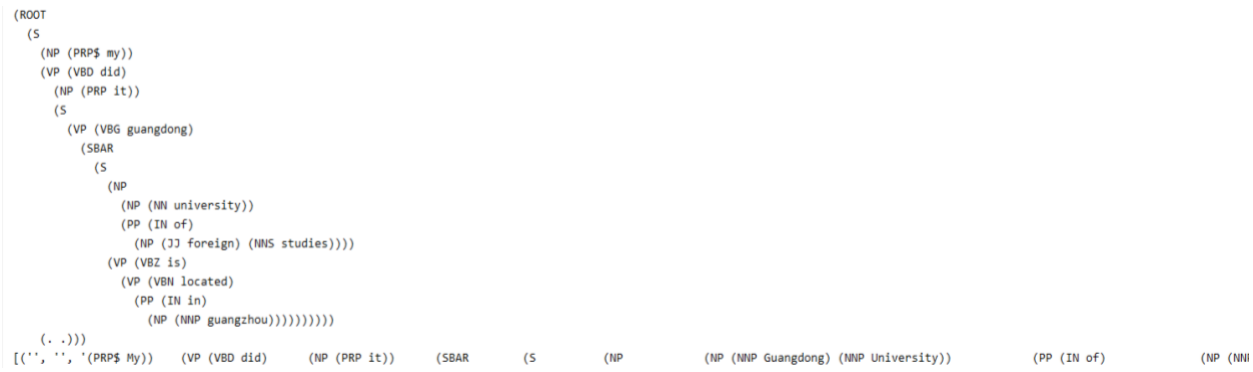

Fig. 8: An Example of Stanford CoreNLP Library to Parse a sentence, IST 664

Lastly, we evaluated the performance of each model and found that the model of past tense detection with Stanford CoreNLP library generated a better result (whose accuracy rate is 72.15%). We also concluded difficulties and limitations in the paper. One difficulty was defining personal

experience or storytelling since the boundary between self-disclosure and personal experience is nebulous. Personal experience focuses on actual events that happened previously, whereas self-disclosure emphasizes emotion sharing. Also, people revealed their own experience to enhance the information reliability while in the area of self-disclosure, people revealed personal information to enhance interpersonal relationships and improve social support. As for limitation, one limitation was the inability of the Stanford CoreNLP library in sentence parsing. Another limitation was that our past tense detection is based on regular expression. It was possible that the verb in past tense failed to connect with subjects (such as I, we, my, our) if the interval between the verb and subjects was big.

IST 700: Deep Learning, NLP, and Computational Social Science

Similarly, IST 700 focuses on the domain of academic language analysis rather than business decision makings. The final project was to generate a COVID-19 Public Sentiment Analysis which captures popular topics and associated sentiment dynamics based on the massive social media posts on Twitter. Our group used Python Tweepy with Twitter API 2.0. to collect tweet posts that started from March 29 to April 30 daily. Hashtags used to match COVID-19 related contents including but not limited to "#cornonavirus", "#coronavirusoutbreak", "#covid", "#covid19", and "#ihavecorona". 14,607,045 tweets (5.31G) had been downloaded, containing contents posted in multiple countries and various languages. After location and language filtering, only 161,220 English-written tweets originating exclusively in the United States remained. It is necessary to point out that tweets without specifying location have not been taken into account. Three out of the four groupmates were required to manually annotate 100 sample data extracted from the April collection, reaching an agreement with Cohen's Kappa value of 0.665.

After data acquisition, several data processing techniques, including emoji removing, stop-words filtering, and POS tagging matching, were executed, following by model construction and topic modeling. For topic modeling, an LDA using Java-based Mallet software package was utilized to identify the most popular COVID-19 related topics in New York State and California. In the process of model development, five models have been constructed, including fasttext, Random Forest, XGBoost, LSTM, and pre-trained BERT. When it comes to model evaluation, BERT with an accuracy rate of 0.702 outperformed the others in terms of accuracy, recall, precision, and F1 score.

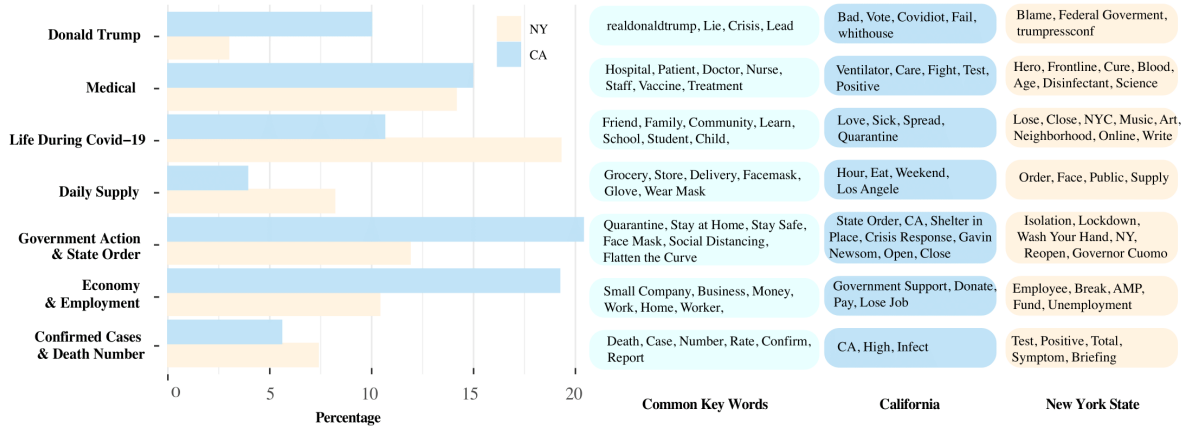

Fig. 9: Themes (by %) mentioned in CA and NY, with the top associated words, IST 700

Based on the probabilities included in the document list, 35 topics were generalized into seven categories after merging similar topics and eliminating irrelevant topics (Fig. 9). For New York States, 19.28% of the contents pertained to the theme of Life During COVID-19, followed by the approbation toward medical staff and concerns toward disease (14.14%), discussion on state policy (11.88%), and the impact of COVID-19 on the economy and employment (10.39%). Compare with New York, people in California showed more concern on Policy related topics (which is 20.38%), followed by the impact on economy and employment (19.22%), and hospital and medical issues (14.95%).

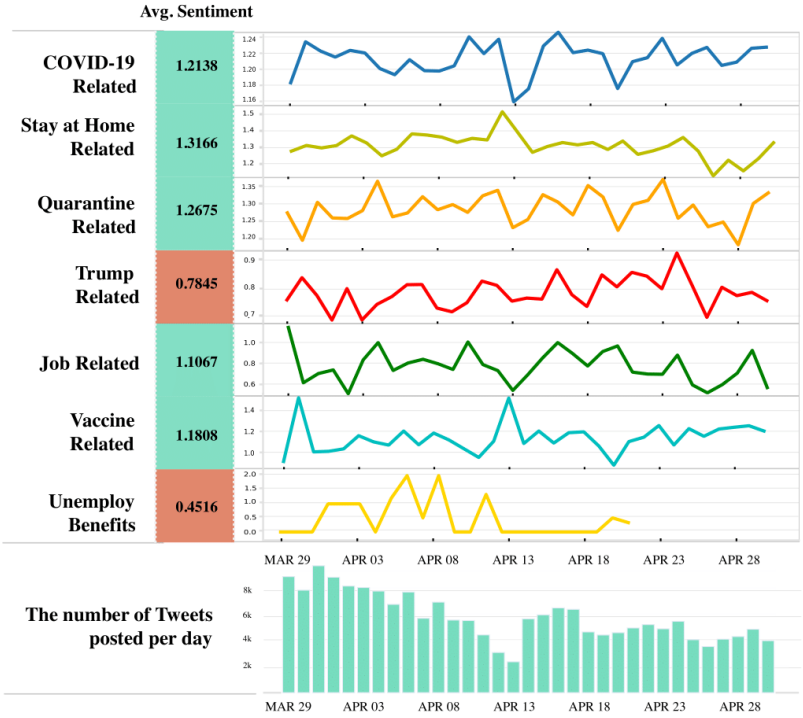

Fig. 10: Topic level Sentiment Changes, IST 700

Based on the results of LDA modeling and the ranking result of hashtags with a high-frequency value, the top seven hashtags are selected, including COVID related tags (e.g., covid, coronavirus), trump, quarantine, job-related tags (e.g., unemployment, work), vaccine, stay at home, and unemployment benefits. In the next step, we used rules and regex to find the group each tweet belongs to, and then to predict sentiment score. Interestingly, scores related to Donald Trump are lower than any other frequent topics, except for "Unemployment Benefit" whose low value was caused by an inadequate amount of data. Moreover, people hold a relatively positive attitude toward Stay-at- Home policy compared with the Quarantine policy. One speculation is that people tend to espouse a voluntary order instead of a mandatory action.

Through this project, we found that different states showed disparate interest and emotion towards the covid19 situation. People in California pay more attention to government action and economy while people in New York focus on daily life during COVID-19 and Medical related events. What's more, people showed more concern for the pandemic's impact on their personal life (such as losing jobs and unemployment benefits). The unemployment trend during COVID-19 outbreaks stress and bring a huge negative emotion to people on Twitter, and the unemployment benefits also give people dramatic emotional change in April. Through examination of sentiment changes and trends with related topics and themes, the government agency, health organization, the business industry could gain an informative understanding of addressing issues aroused during the pandemic period. And the way to use transfer methods predicts social media sentiment can provide an efficient and urgent view of different groups people's emotions.

Conclusion and Reflection

To sum up, I completed various projects and research on different fields and topics during the 2-year master program in iSchool. From a technical perspective, this program advanced my ability in statistics, math, and computer science. Besides, all the team projects I participated in cultivated my soft ability in communication, collaboration, and presentation. As a qualified graduate student, I fully testified and exemplified my ability in data processing and management in all these data science projects. After graduation, regardless of conducting an individual assignment or being a team member, I am being able to finish tasks included in a data science process:

- Define and understand business questions and business goal
- Collaborate and communicate with others if required

- Data acquisition and processing, including data preprocessing and feature engineering
- Data understanding and visualization
- Model development, evaluation, and selection
- Software deployment if applicable
- Interpretate model result and inference
- Generate business insights and make business decisions
- Awareness of ethical issues in data science practice

Being one graduate student of the ADS program is an unforgettable and precious experience. Not only have I learned how to work agilely as part of the development team and lead a small group of people from different background to collaborate and work towards a common goal, but also have gained hands-on experience through my academic projects. Therefore, I am confident that I am well-prepared both academically and practically.