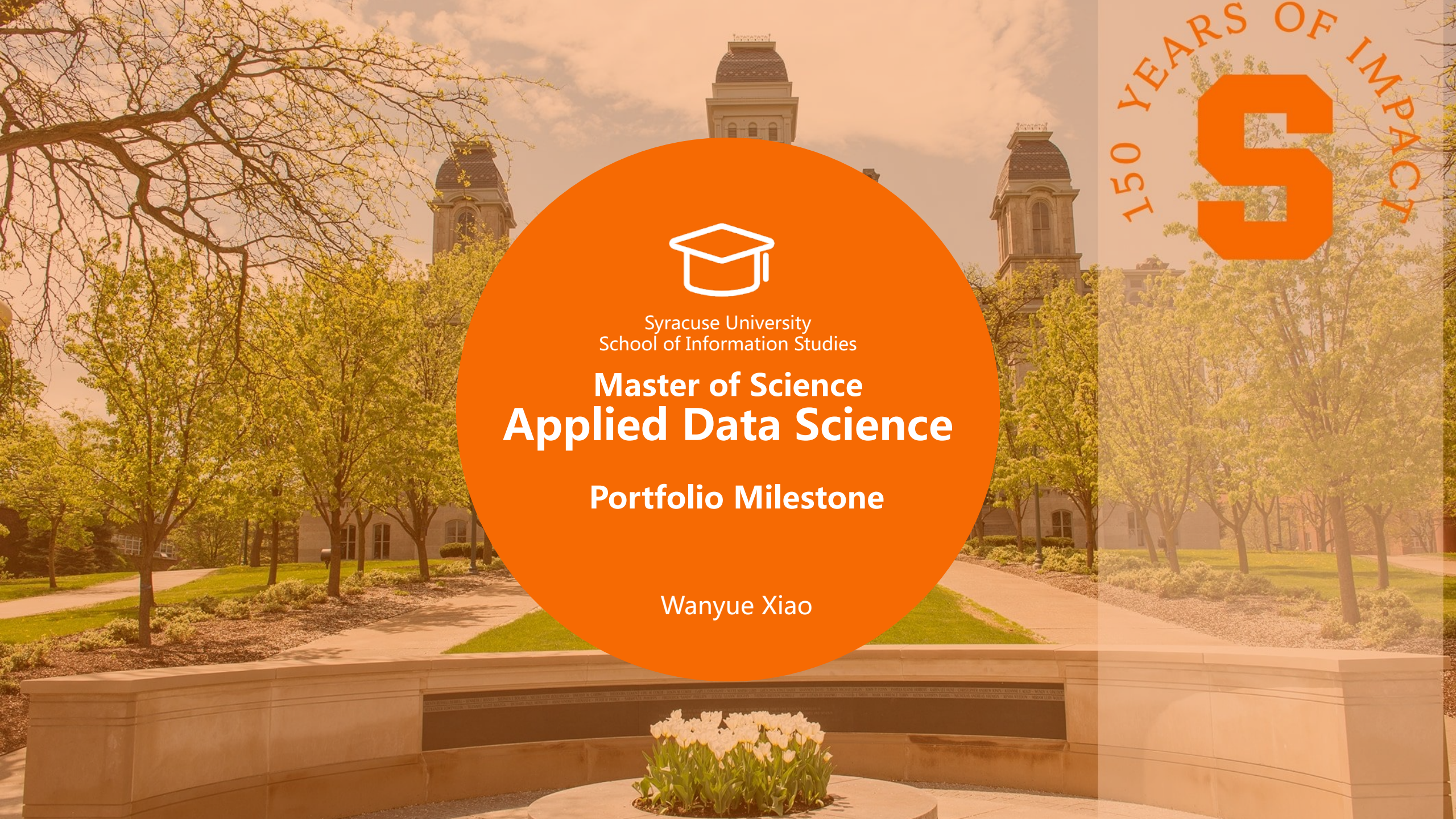


Syracuse University
School of Information Studies

Master of Science
Applied Data Science

Portfolio Milestone

Wanyue Xiao

Learning Outcomes

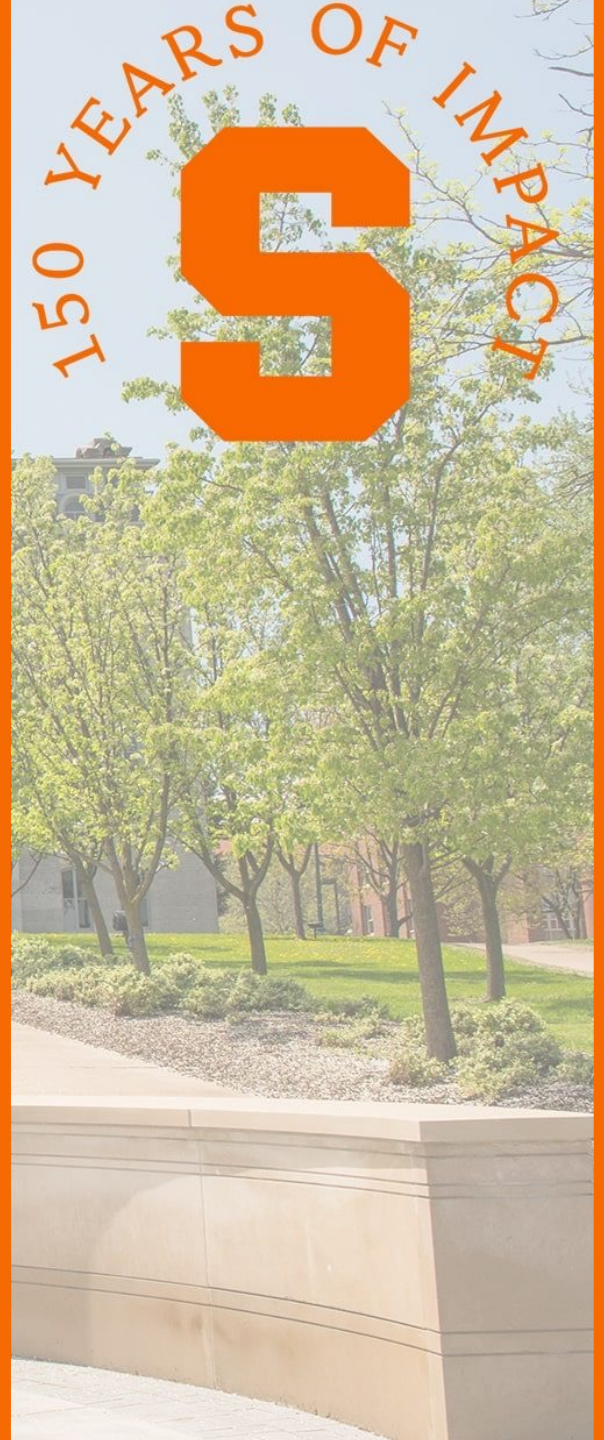
Applied Data Science

Technical Skills:

- Describe a broad overview of the major practice areas of data science
- Data acquisition and organization
- Identify patterns through visualization, statistical analysis, and data mining
- Develop alternative strategies based on the data
- Develop a plan of action to implement the business decisions derived from the analysis
- Synchronize the ethical dimensions of data science practice

Soft Skills:

- Develop communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals
- Demonstrate collaboration skills with colleagues



Academic Projects

From 2019 Fall to 2020 Fall

2019

2019 Fall

IST 659 Database Administration

ToyPedia Online Forum Database Design

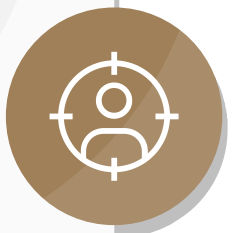
Provided a database solution for an online forum.

2020

2020 Spring

IST 719 Information Visualizationn

Video Games Sales Analysis



Created a poster based on a video game sales dataset.

IST 707 Data Analytics

US Car Accidents Severity Analysis

Unearthed the causes of U.S.A. vehicles accidents and predicted the severity level based on data.

2020
2020 Spring

IST 718 Big Data Analytics

Telecom Customer Churn Analysis

Discovered critical features influencing customer churn and proposed business actions based on the analysis.

2020
2020 Fall

IST 700 Deep Learning, NLP, and Computational Social Science

COVID-19 Sentiment Analysis

Found topics of COVID-19 tweets posted in 2020 April and developed model to classify users' emotion.

2020
2020 Fall

IST 659

Data Admin & Database Management

Programming Language:

- SQL

Tools or Skills:

- Microsoft SQL server
- Visio
- Data Modeling & Design

Read more

Project Description

Introduction

Under the instruction of Prof. Hernando Hoyos, our team re-designed a database for an online web-based forum called ToyPedia where toy collectors shared information of their collections with others.

The relational data model should be divided into four parts, including

- **External User Section** which defines external users who might access data or have influence of data flow;
- **Posting and Replying section** which allows users to post, reply, rate, subscribe content;
- **Purchasing Section** which enables users to purchase products;
- **Payment and Delivery Section** which stores detailed information on payment method, transaction data, and delivery records.

1

Database Incompatibility

The development team failed to restore data on the new version of the existing database.

2

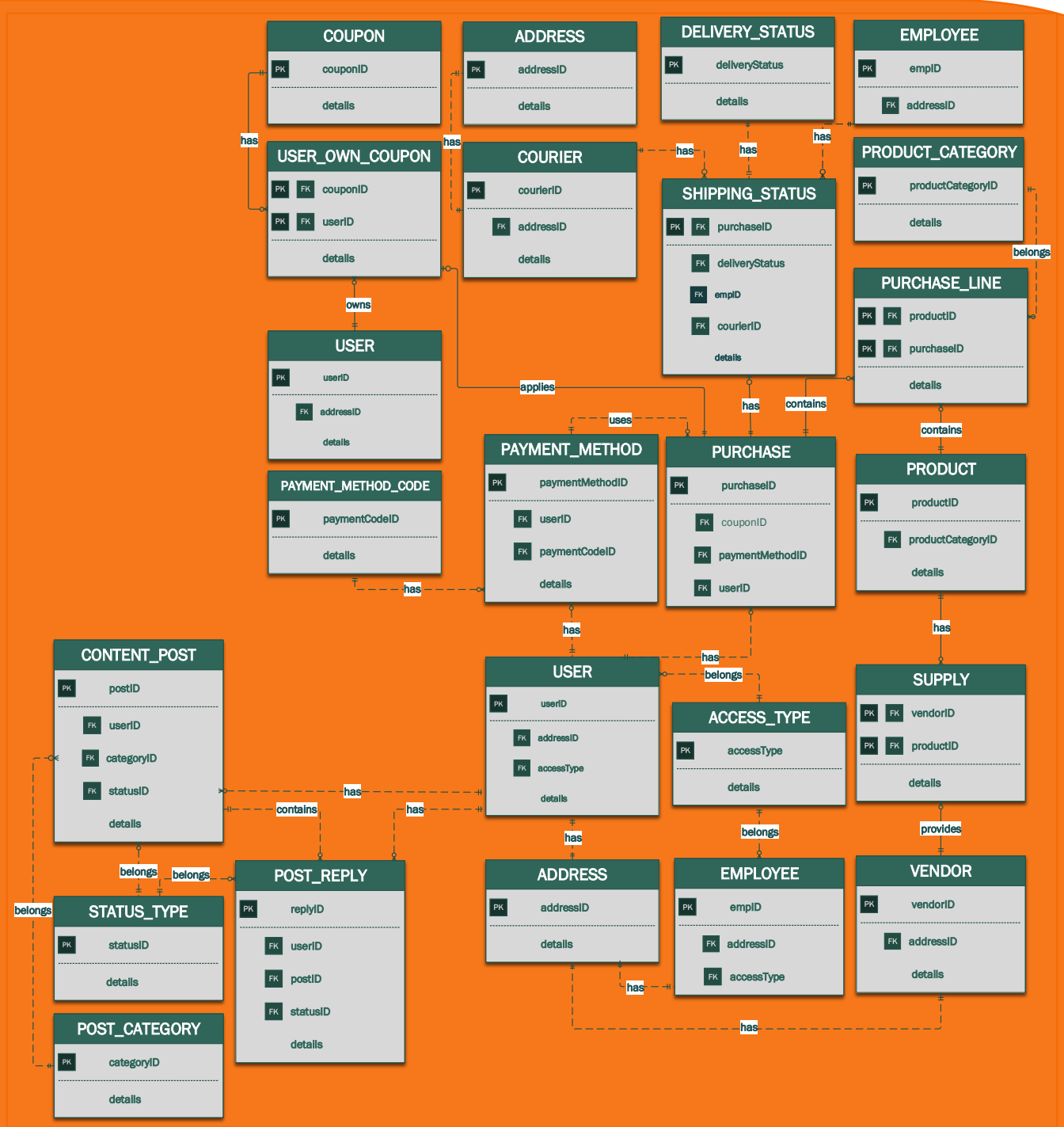
Limited Storage Capacity

With the growth of users, it put pressure on the existed database.

3

Data Inconsistency

Some entities had been deleted, yet their primary keys were kept in the other entities.



Business Rules

1) Each employee has a unique discount rate when they purchase an item.

2) The delivery fee of an order should be free if the order is above $20.

3) The final rating of a post should be the average of all the ratings given by many users.

4) A product can be supplied by many vendors and a vendor can supply multiple products.

5) A user must provide his or her credit/debit card details to purchase a product.

Which users have purchased products from the forum previously?

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

Data Questions

1) Which product is sold the most by the forum?
2) Which user buys the greatest number of products from the forum?
3) Which product has the highest revenue?
4) What is the average number of orders per user?
5) What is the annual revenue of each product supplied by the forum?
6) Which city has the greatest number of orders?
7) Which user's post gets the highest rating?
8) Which product is sold the least by the forum so that the forum can discontinue the product?

Reflection and Limitation

Develop a database management solution from scratch

This project provides an opportunity for me to develop a database management solution from scratch, including user requirement analysis, business rules affirmation, ERD construction, database development, and prototype development.

Opportunity to practice data management skills

Through this project, I have applied all the data management skills (data retrieving, data integration, and data normalization in relational database) to a real business process.

Enhance database's effectiveness and efficiency

By setting up stored process, report, and trigger, redundant workload could be avoided.

IST 707

Data Analytics

Programming Language:

- Python, R

Tools or Skills:

- EDA
- Statistical Analysis
- In-depth Machine Learning
- Introductory Deep Learning

Read more

Project Description

Introduction

To find the causes of car accidents, our team analyzed the US Accidents dataset using data mining methods and created different models to predict the severity of accidents.

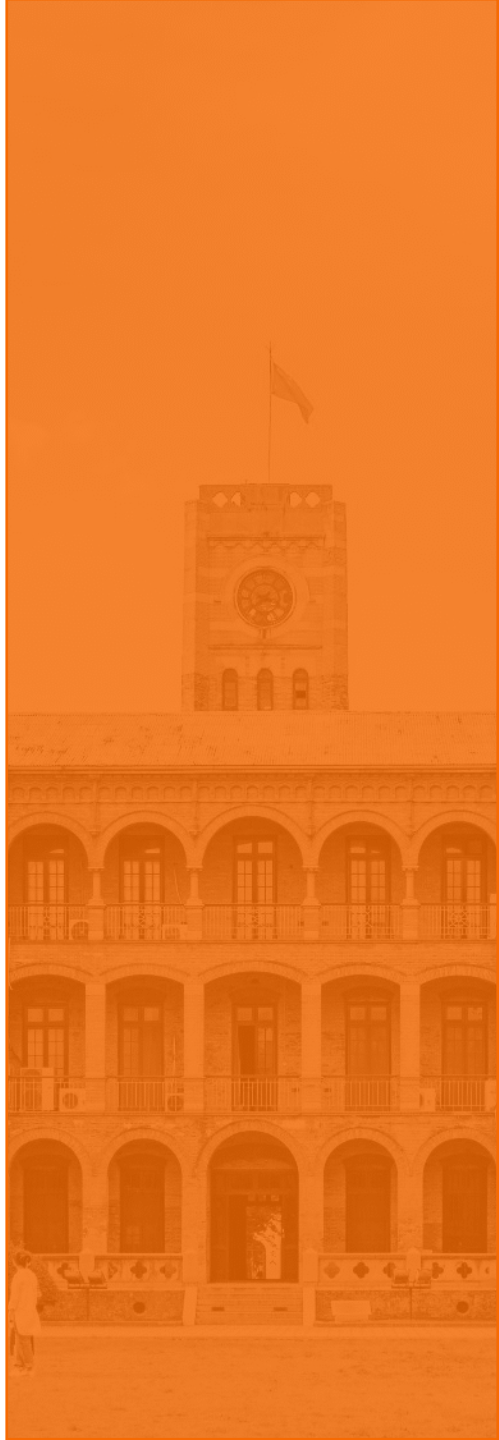
The process of data mining could be separated into two groups:

- Descriptive approaches
- Predictive approaches

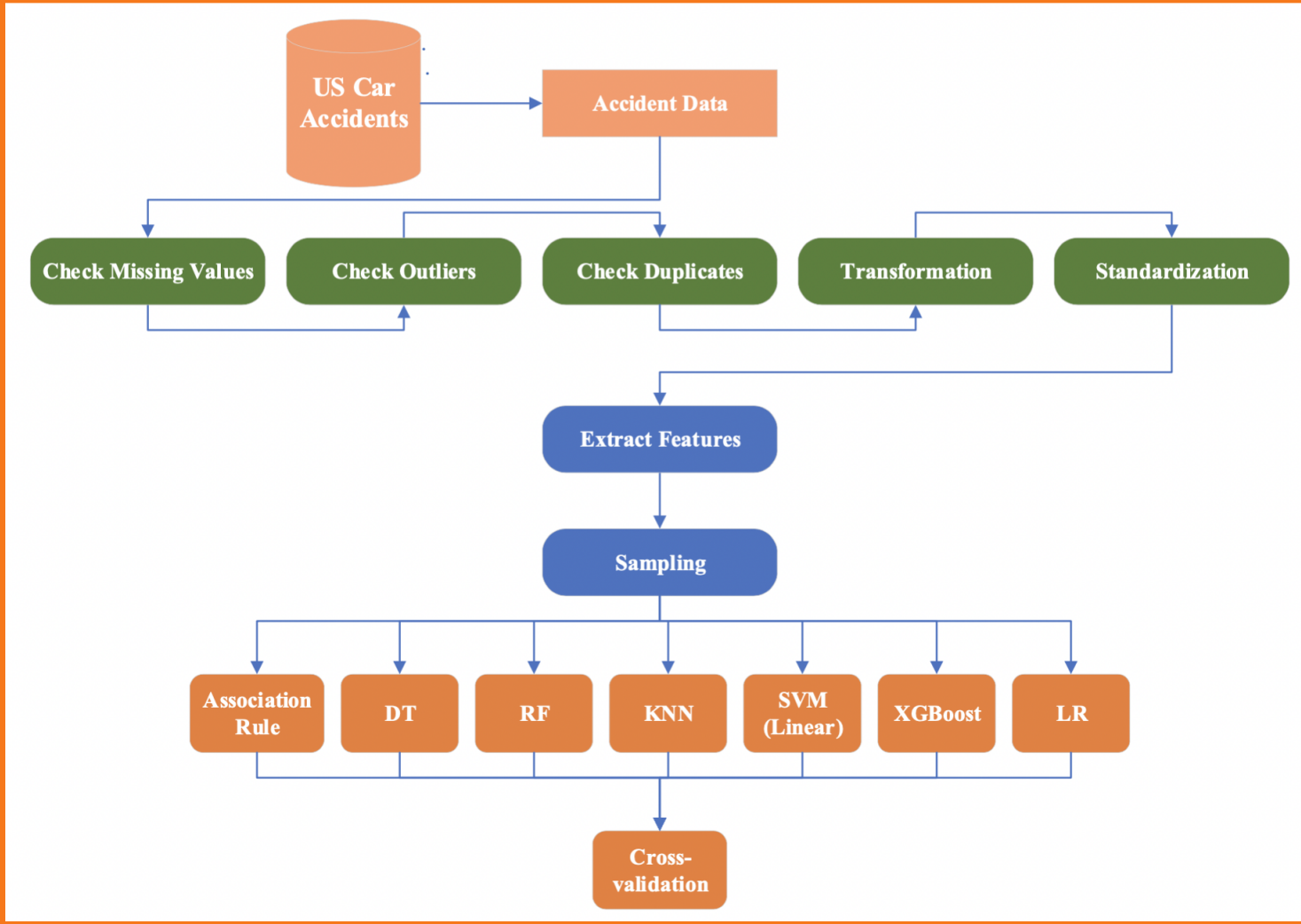
Dataset Description

The dataset is about car accidents that happened in 49 states of the United States from February 2016 to December 2019, containing 3.0 million records in total.

- Randomly sampled and downloaded 50 thousand records and 49 variables
- The accident severity was labeled as Low (severity 1 and 2) and High (severity 3 and 4)



Overall Process of Study



Experiment Design

Five components for the process:

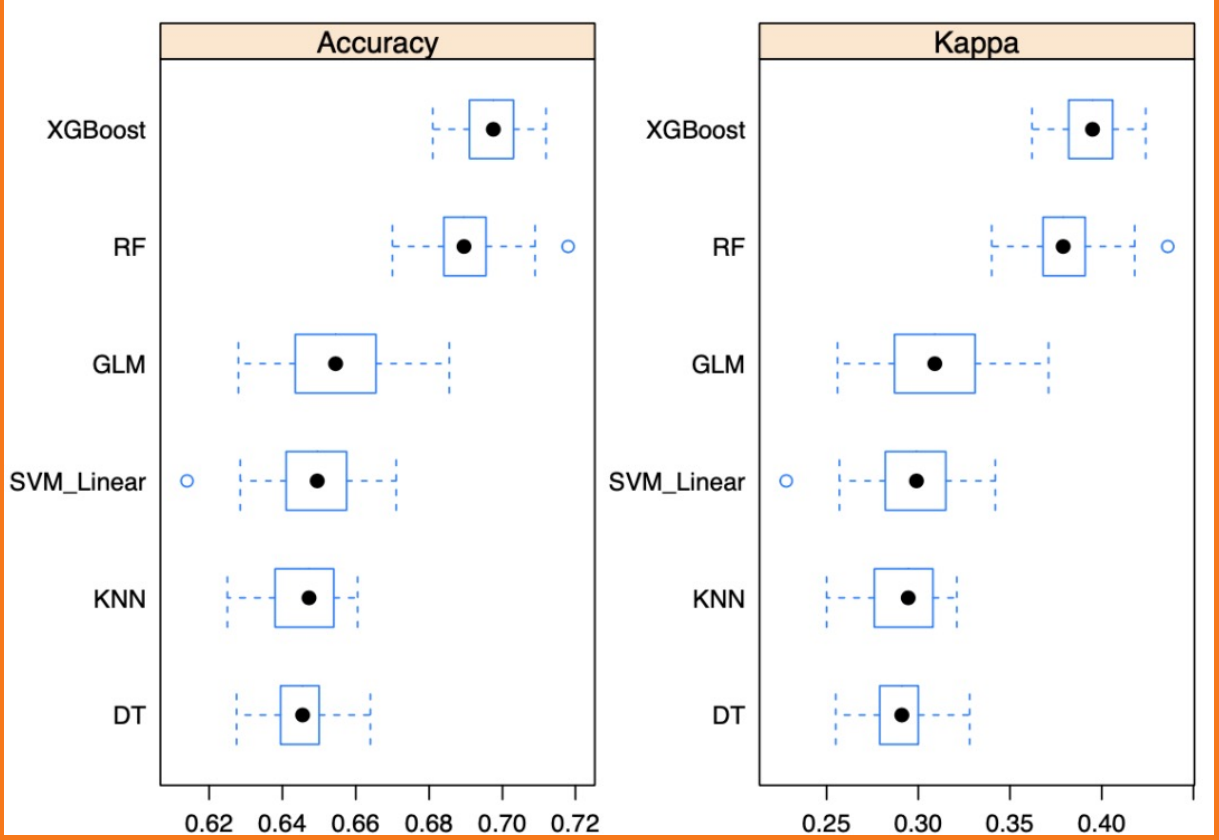
- Data preprocessing
- Exploratory Data Analysis
- Data sampling
- Model generation
- Performance evaluation

Three types of machine learning algorithms used in this analysis:

- Association Rule Mining
- Classification models including Decision Tree, Random Forest, K Nearest Neighbor, Support Vector Machine, Gradient Boosting Machine
- Regression model such as Logistic Regression

Accuracy Comparison of Six Classifiers with 10-fold Cross-validation

	XGB	RF	GLM	SVM	DT	KNN
Accuracy	0.6894000	0.6776000	0.6560000	0.6498000	0.6418000	0.6416000
Sensitivity	0.8012000	0.7852000	0.6924000	0.7964000	0.7724000	0.6820000
Specificity	0.5776000	0.5700000	0.6196000	0.5032000	0.5112000	0.6012000
Precision	0.6547891	0.6461488	0.6454139	0.6158367	0.6124326	0.6310141
Recall	0.8012000	0.7852000	0.6924000	0.7964000	0.7724000	0.6820000
AUC	0.7662403	0.7421255	0.7256226	NA	0.6865785	0.7025220



Summarization

We analyzed the predictive power of six classifiers which were tested on the training dataset by using the 10-fold cross validation.

The observation suggested that:

1. The optimized XGBoost provides the best performance.
2. The RF is the classifier that has the second-highest accuracy rate and second-highest precision value.
3. The DT model performs worst in this task.
4. The algorithms of GLM and KNN show acceptable performance with an AUC value larger than 0.7 respectively.

Reflection and Limitation

Exploratory Data Analysis is essential

Having EDA help data scientist and data analyst understanding data before making any assumptions. It help professionals identify obvious errors as well as better understand patterns within the data.

Taking computation costs and running time into consideration

The computation cost of certain algorithm could be significantly high. We should try to find a balance between precision and accuracy. Additionally, one can use alternative strategies to reduce running time and computation cost.

Comparison between R and Python

R is capable of plotting beautiful plots, whereas Python is capable of managing data in large volumes.

IST 718

Big Data **Analytics**

Programming Language:

- Python

Tools or Skills:

- Apache Spark
- EDA
- Machine Learning & Statistical Learning
- Time Series Analysis

Read more

Project Description

Introduction

This study aims to discover critical features that influenced customer's churn rate based on a telecom customer dataset.
Our team developed several models that are capable of predicting customer churn after comprehensive data exploration and analysis.

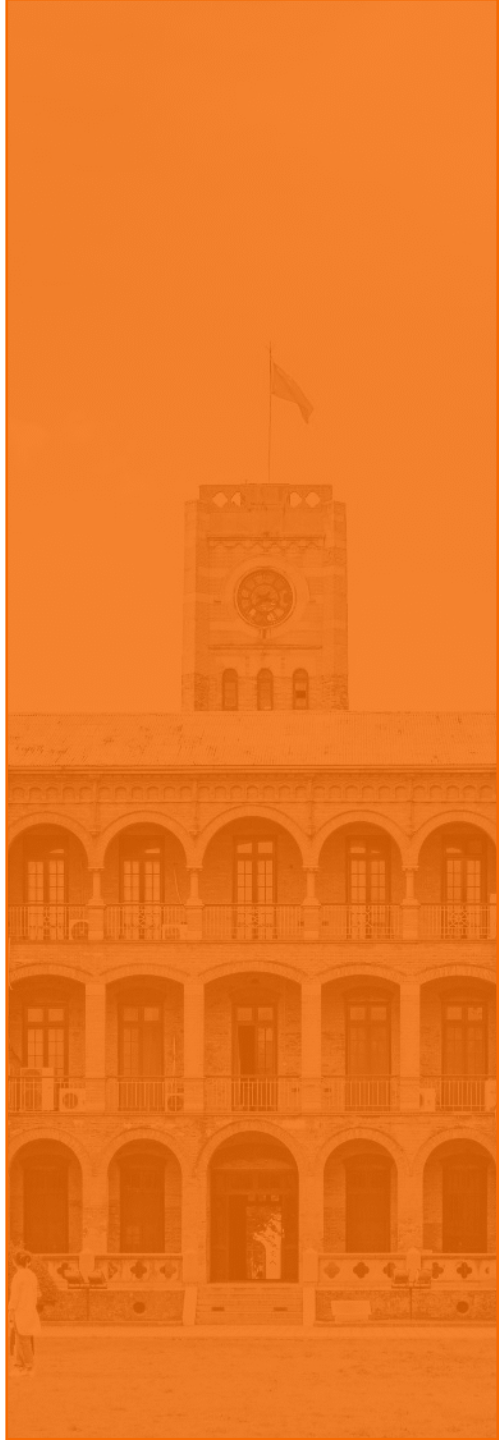
This data science project could be separated into two groups:

- Descriptive approaches, such as EDA
- Predictive approaches

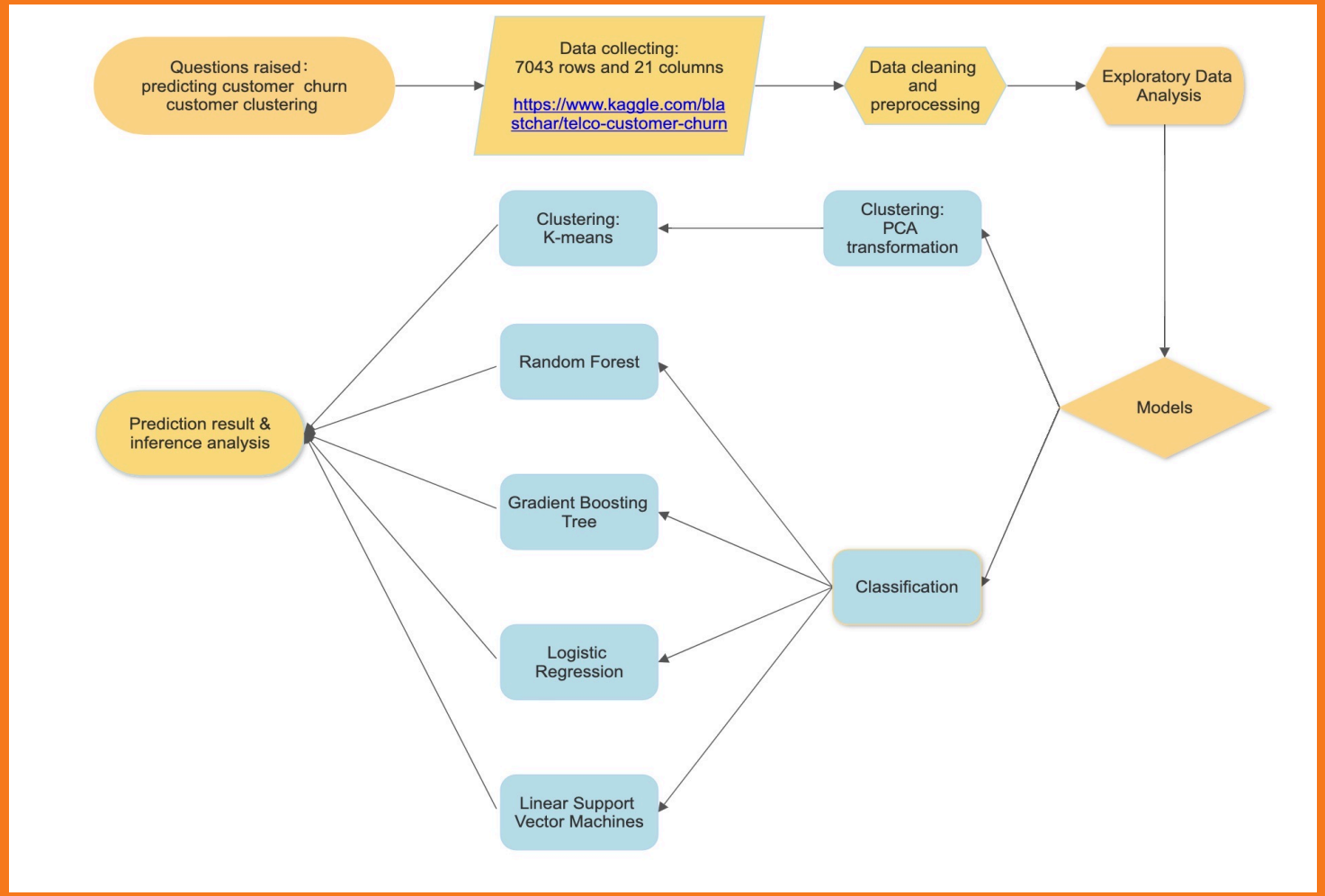
Dataset Description

The telecom customer churn dataset could be separated into three categories, personal attributes, accounts details, and services registered up by the customer.

- The dataset contains 7043 rows and 21 columns.
- The accident severity was labeled as Low (severity 1 and 2) and High (severity 3 and 4)



Overall Process of Study



Experiment Design

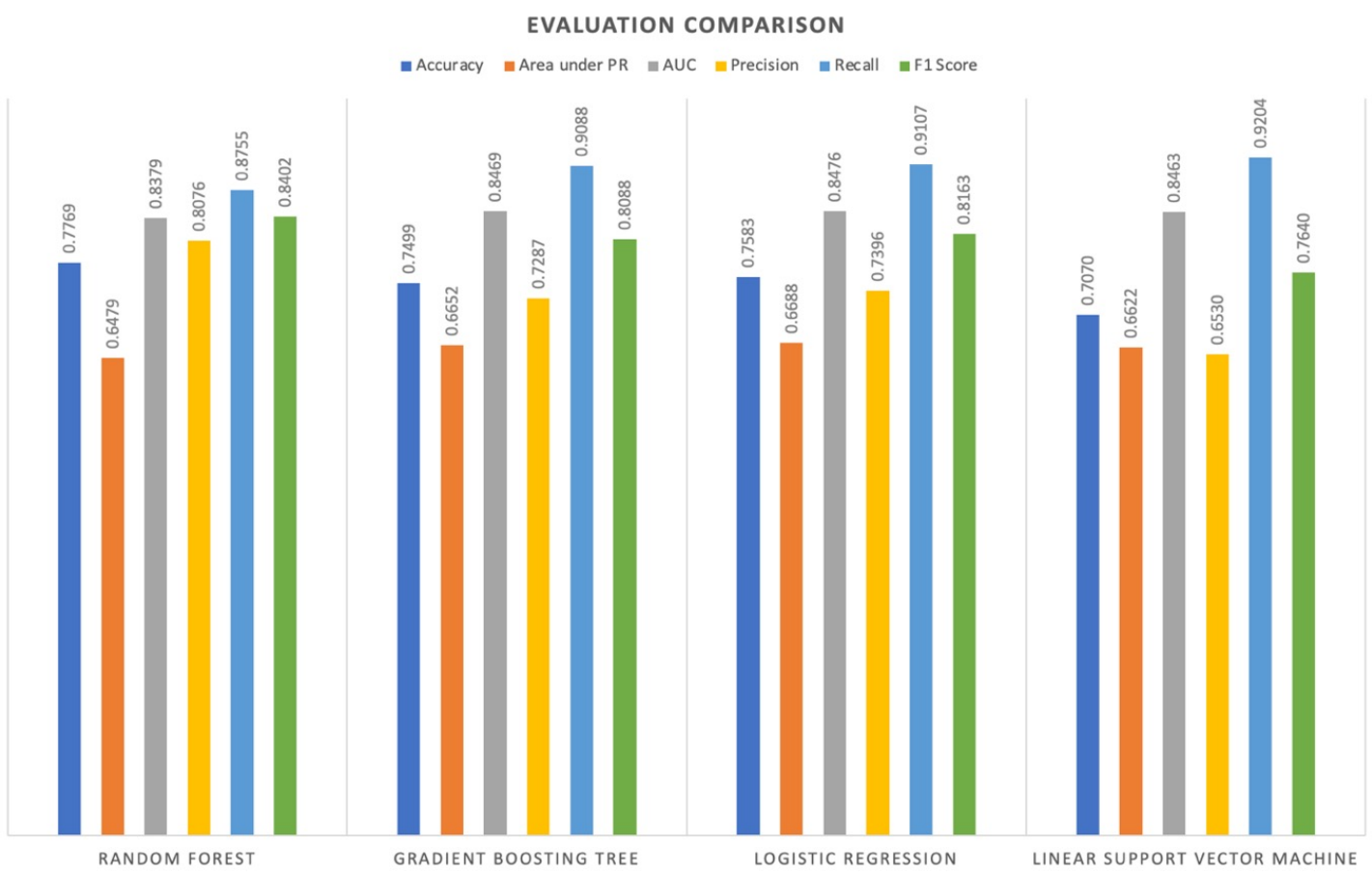
Four components for the process:

- Data preprocessing
- Exploratory Data Analysis
- Model generation
- Performance evaluation

Four types of machine learning algorithms used in this analysis:

- Principle Component Analysis,
- Clustering model, such as k-means
- Classification models including Random Forest, Gradient Boosting Tree, Linear Support Vector Machine
- Regression model, such as Logistic Regression

Accuracy Comparison of Four Classifiers with 3-fold Cross-validation



Summarization

We analyzed the predictive power of 4 classifiers which were tested on the training dataset by using the 3-fold cross validation.

The observation suggested that:

1. An average accuracy of 0.74 could be obtained across the four supervised models to predict churning customers.
2. The Random Forest model outperformed the others in the accuracy rate (0.7769) and F-measure (0.8402).
3. Logistic Regression and Gradient Boosting have more strength in minimizing the false negatives.
4. The observations suggested that a linear model is capable of describing the data relationship.

Reflection and Limitation

Dealing with real-world business problem

This project proffered an opportunity for me to generate busines insights from a real-world business problem. Through organizing, analyzing, and visualizing the data of customer transactions and customer information, we found patterns for customers who stick to the company and customers who tend to leave.

Having prediction and inferences in advanced

Writing down predictions and inferences before analysis is beneficial, providing guideline for the organization of data science project.

Staying open-minded during analysis

Through this project, I have learned that having an open-minded attitude is critical during a data science process. However, we shall not accept patterns derived from data mining analysis blindly given that we are responsible to interpretate results to our clients.

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

Programming Language:

- Python

Tools or Skills:

- Data mining methods,
- Topic Modeling,
- Corpus Annotation,
- Word Embedding
- Transfer Learning

Read more

Project Description

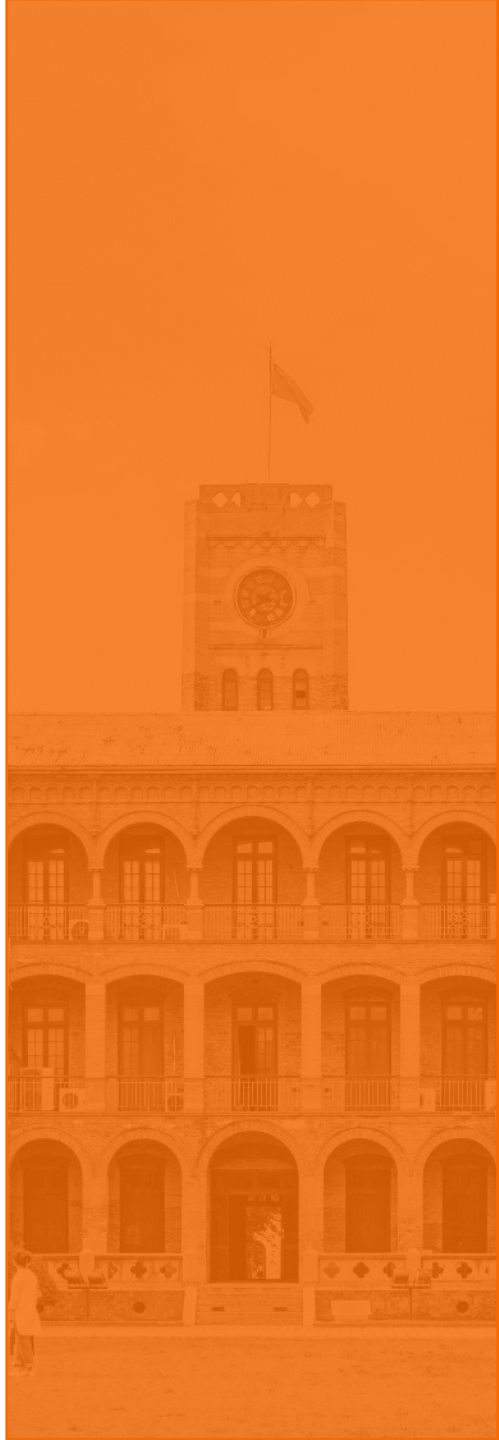
Introduction

This project is to conduct a COVID-19 public sentiment analysis which captures popular topics and associated sentiment dynamics based on the massive social media posts on Twitter.

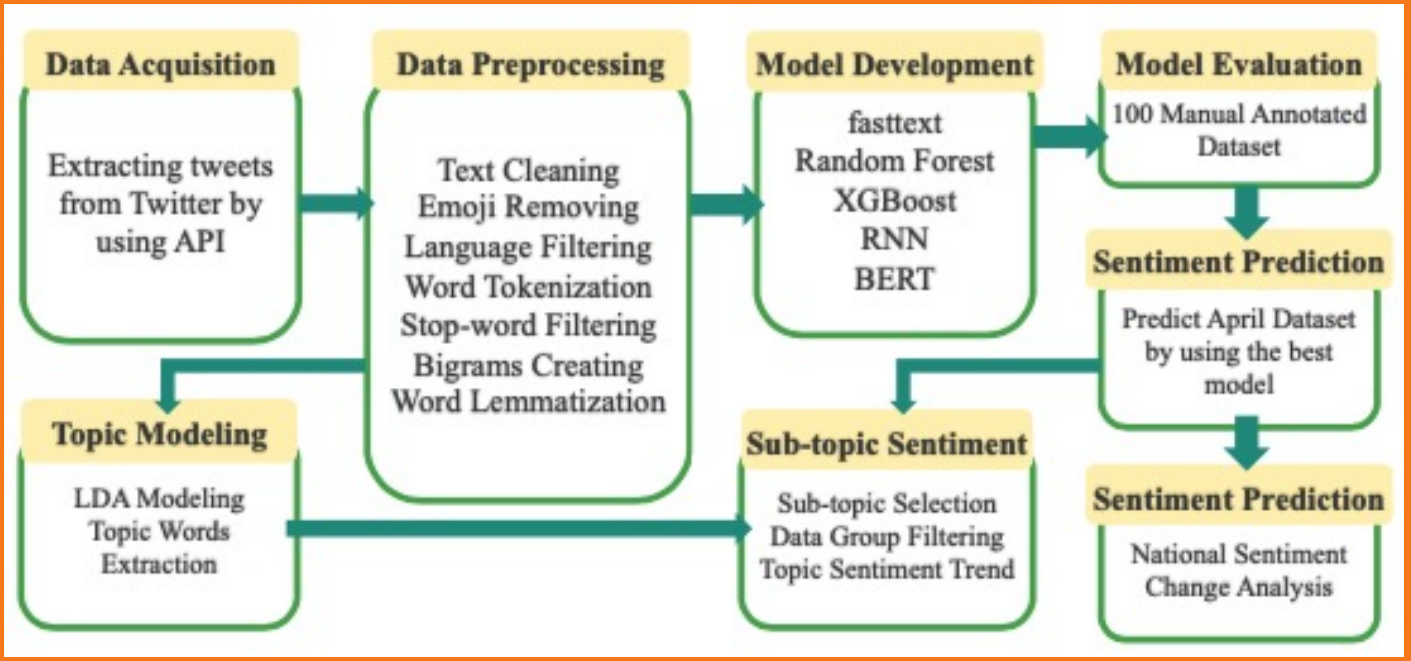
Dataset Description

Our group used Python Tweepy with Twitter API 2.0. to collect tweet posts that started from March 29, 2020 to April 30, 2020.

- 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.
- Only 161,220 English-written tweets originating exclusively in the United States remained.
- Tweets without specifying location have not been taken into account.



Overall Process of Study



Model Evaluation

Model Name	Accuracy	Precision	Recall	F1
fastText	NA	0.5677	0.5677	NA
Random Forest	0.610	0.60	0.589	0.581
XGBoost	0.577	0.583	0.577	0.575
RNN	0.6453	0.6399	0.6275	0.6246
BERT	0.7021	0.714	0.691	0.6811

Experiment Design

Five components for the process:

- Data Acquisition
- Data preprocessing
- Model Development
- Model Evaluation
- Topic Modeling

Six types of machine learning algorithms used in this analysis:

- LDA Modeling
- Fasttext, XGBoost, Random Forest, RNN, BERT

Model Comparison:

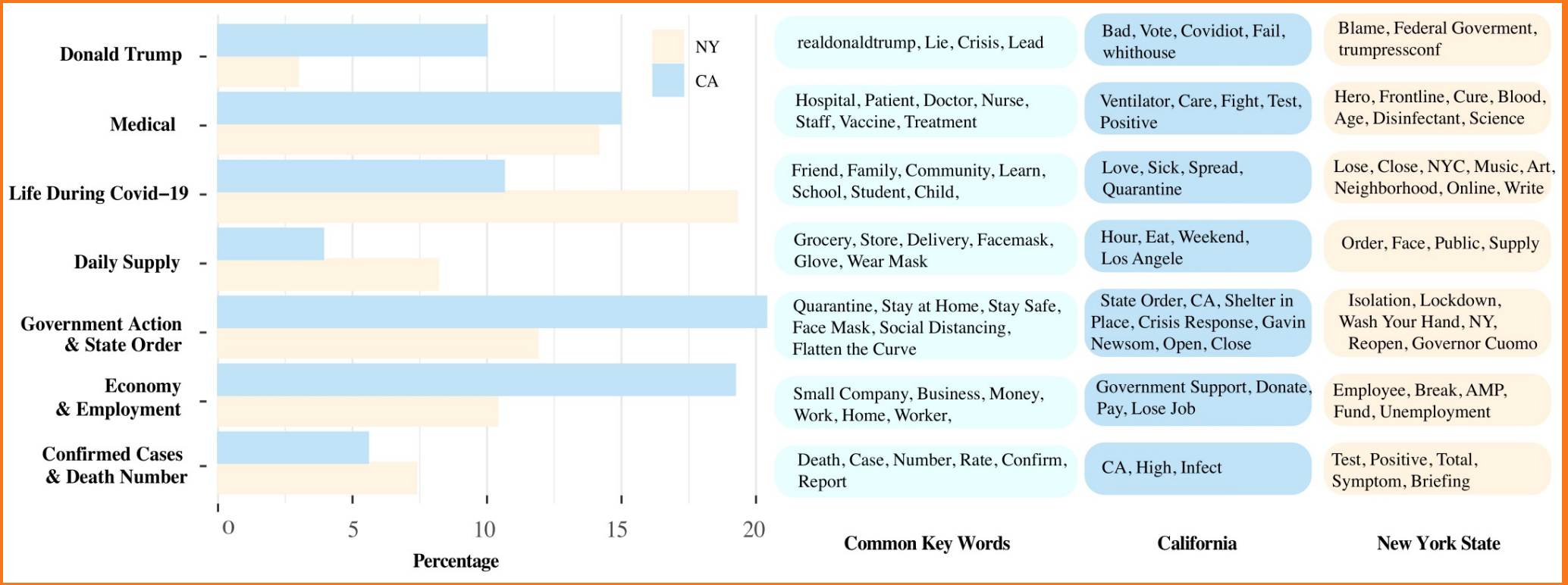
1. The performance of fasttext on the sample data is unsatisfactory
2. BERT with 0.702 in accuracy and 0.681 in F1 score outperformed the others.

Summarization

The observation suggested that:
Citizens in 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.
- People showed more concern for the pandemic's impact on their personal life (such as losing jobs and unemployment benefits)

Result of Topic Modeling



Reflection and Limitation

Opportunity to conduct an academic research project

Unlike profit-driven business project, conducting an academic research project requires rigorousness, bias-free attitude, and validation from ground truth.

Error analysis could be helpful

Error analysis help researcher to reasons of failure after digging into the results of the models after each iteration.

Data Privacy and API Constrains

To protect user's privacy, data collectedd through API should not contain any user sensitive information. Also, getting consent from user also is essential.

Reference List

1. Xiao, WY. (2019), IST 659: Data Admin & Database Management. Retrieved from https://github.com/xwanyue0221/Graduate-Portfolio/tree/main/IST659

2. Xiao, WY. (2020), IST 707: Data Analytics. Retrieved from https://github.com/xwanyue0221/Graduate-Portfolio/tree/main/IST707/Final%20Project

3. Xiao, WY. (2020), IST 718: Big Data Analytics. Retrieved from https://github.com/xwanyue0221/Graduate-Portfolio/tree/main/IST718/Final%20Project

4. Xiao, WY. (2020), IST 700: Deep Learning, NLP, & Computational Social Science. Retrieved from https://github.com/xwanyue0221/Graduate-Portfolio/tree/main/IST700/Final%20Project

Syracuse University
School of Information Studies

Thank You