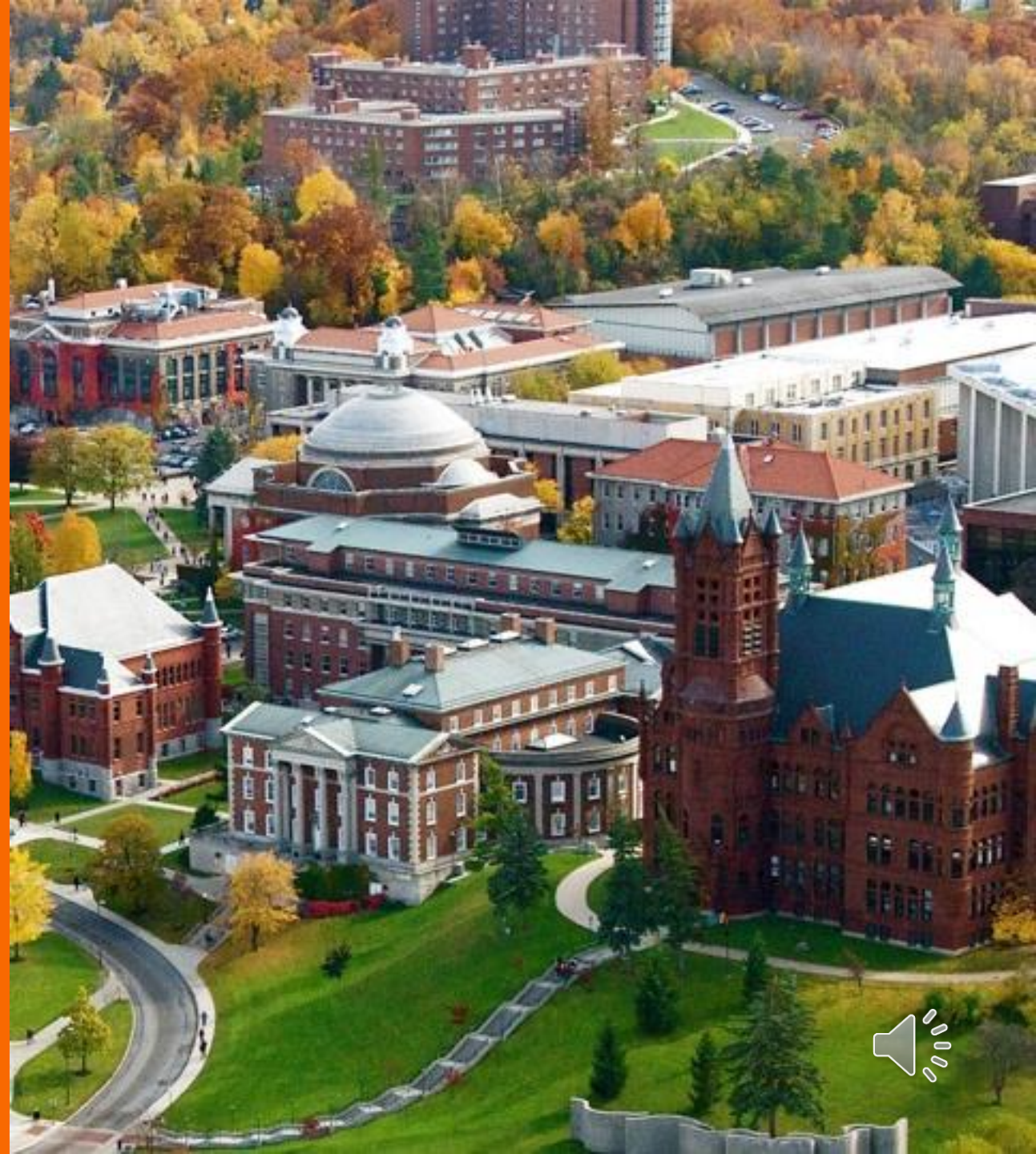
Syracuse University

Master of Science
Applied Data Science

Portfolio Milestone

Ruiwei Zhang
486964430



Introduction

The Applied Data Science program at Syracuse University's School of Information Studies provides students the opportunity to collect, manage, analyze, and develop insights using data from a multitude of domains using various tools and techniques. In courses such as Database Administration (IST 659, 2019), Introduction to Data Science (IST 687, 2019), Natural Language Processing (IST 664, 2020), and Big Data Analytics (IST 718, 2018), reports and presentations were developed to deliver insights using Microsoft Access, SQL Server Management Studio, Python, R and Excel. The skills developed at the School of Information Studies furnish data scientists focused in the field of marketing analytics with the ability to generate value within their organizations and produce actionable recommendations.

The Applied Data Science Program has seven learning objectives which were exemplified by the applications in this portfolio:

1. Describe a broad overview of the major practice areas in data science.
2. Collect and organize data.
3. Identify patterns in data via visualization, statistical analysis, and data mining.
4. Develop alternative strategies based on the data.
5. Develop a plan of action to implement the business decisions derived from the analyses.
6. Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice

Syracuse University

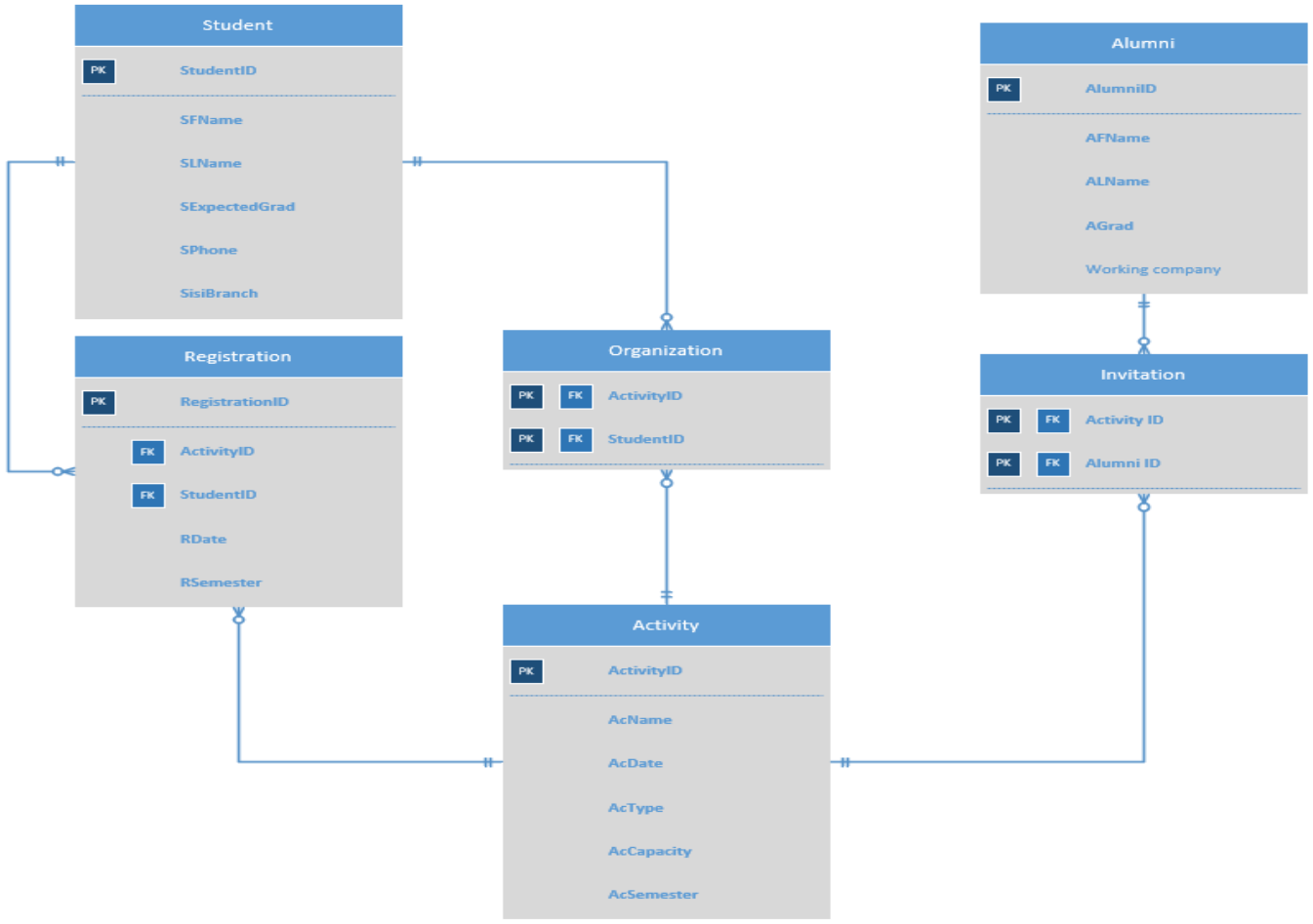
IST-659: IBRANCH NETWORK MANAGEMENT DATABASE

Summary

The Project mainly focus on designing a database system for iBranch, which is a student organization in the School of Information Studies at Syracuse university.

Our proposed system is a method to automate the process of scheduling and reduce the manual work effort for recording attendance. Especially every events site has maximum accommodation, so that it is troublesome for keep calculate current attendant students and reminds how many seats we left. In our system, we can automatically provide current attendant people to organizer, and send a message to potential students to participate in activities, which remind them to sign up to participate form as soon as possible. Also, it is way easier for generate summary report for organization manager, such as total number of attendants this semester, and average attendance rate every month, rather than calculate one by one by hand.

Entity Relationship Diagram



Reflection

- 1. MS Visio: We created entity relationship diagram using MS Visio. It shows the complete structure and the relationship of our database.
- 2. SQL Server: We used SQL Server as the database that stored all the tables and their data. We use several queries to create, insert and select the correlated data set and data.
- 3. MS Access: We used MS Access to create the interface for the system. Using Access, we link our tables that were created in SQL Server. Once, the tables were linked, we created forms that could take user input or display the necessary information to the users. Based on the data, we used MS Access to generate reports for the users of the system.

Syracuse University

IST-687 PROJECT

Introduction and descriptive statistics

- Customer churn is a lagging indicator. Customer churning is happening so we need to take action as soon as possible to keep our customers. We need to find the right leading indicators to help us to provide the right services that the customers need in order to keep them flying with Southeast.

	Min	Q1	Median	Mean	Q3	Max
Age	15	33	45	46.32	59	85
Price Sensitivity	0	1	1	1.279	2	4
Loyalty	-0.9762	-0.7037	-0.4410	-0.2766	0.0588	1
Total Freq Flyer Accts	0	0	0	0.8928	2	10
Likelihood to Recommend	1	6	8	7.073	9	10

- From the above figure, we find that our customers are not price-sensitive. The mean loyalty is negative, providing some insight that customers may not be as loyal as we think. It also suggests that most of the customers do not really care about the loyalty program. Lastly, the average likelihood to recommend our service seems to be not too bad.

Methodology

- Data cleanse
  - NA cleaning: In the data cleaning part, the NAs should be repaired first.
  - Data split
  - For different models, the requirement for data type is distinct. Hence, we need to transform the data type according to the model requirement.
- Modeling technique
  - Linear model
  - Apriori model
  - SVM

Conclusion

- Based on analysis from our three models, we suggest Southeast Airline to put less emphasis on Current Loyalty Program and pay more attention to the factors that affect the customers' satisfaction the most.
- Partnership with Northwest Business Airlines has a negative effect on overall NPS of Southeast Airline.
- Partnership with FlyFast Airways has decreased total NPS.
- More partnership with the Sigma Airlines Inc should be useful for the overall NPS of Southeast Airline.
- Improve service quality in northern regions, especially during the winter.

Reflection

- This result is an example of the importance of testing different data mining techniques to develop the simplest, most accurate prediction models.
- Testing alternative strategies and weighing the benefits of each technique with respect to the data can reduce computation costs and provide the greatest precision.
- This is an important distinction in a marketing analytics setting, which is magnified by the scale of the data.

Syracuse University

IST-664: BOSTON AIRBNB FEEDBACK ANALYSIS

Abstract

- Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. The Airbnb has been trying to improve their service. In the past few years, applying Natural Language Process (NLP) to analyze users' satisfaction becomes a trend for service improvement. Hence, we find a dataset of the guests' comments on the Airbnb services.
- This dataset describes the listing activity and metrics in Boston. We are focusing on the following Airbnb activity included in this Boston dataset. Based on the comments of the reviewers, we want to dig deeper on the general feedbacks of each Boston neighborhood and give specific suggestions on the Airbnb services.
- The sentiment analysis will be done on both the paragraph and the sentence level. What users pay attention to is the target of the analysis.

Methods

- Our sentimental analysis will be applied on both sentence and comment level.
- The k-means model will be applied to discover the similar patterns. This will classify the words in the comments into different categories.
- The sentiment polarity of comments is also in our consideration. In this step, words will be judged whether they are positive or negative.
- After this step, it could be seen clearly that what kinds of words are the guests' focus. In addition, the Latent Dirichlet allocation (LDA) will be used for topic analysis. It can separate the topics that users pay most attention on. In this way, specific suggestions can be given to the Airbnb for future improvement.

Conclusion

- After summarizing the results got from all three models, we can make a conclusion for our business question: what are the major factors that affect customers' satisfaction.
- First, cleanliness. Customers really care about if the room, bathroom, the sheet on the bed is clean or not. It is very important factor for them to rating a house or apartment.
- Second, facility, like bed is comfortable, the kitchen is well equipped, etc.
- Third, location. The house near to public transportation and grocery store are always very popular. Because customers can easy travel to other place for their trip and they can buy necessary supplies nearby.
- The last one is about service. Lots of comments mentioned about great host , feel at home. So we think the service provide by host, or the attitude of host may be a significant factor. A nice host will leave great impression to customers.

Reflection

- This exercise provided the opportunity for the collection and structuring of externally-sourced data, identification of patterns within and between clusters of text, and developed insights into the behavior of elected officials.
- User privacy was considered to both request only the necessary data and maximize the API rate limitations.

Syracuse University

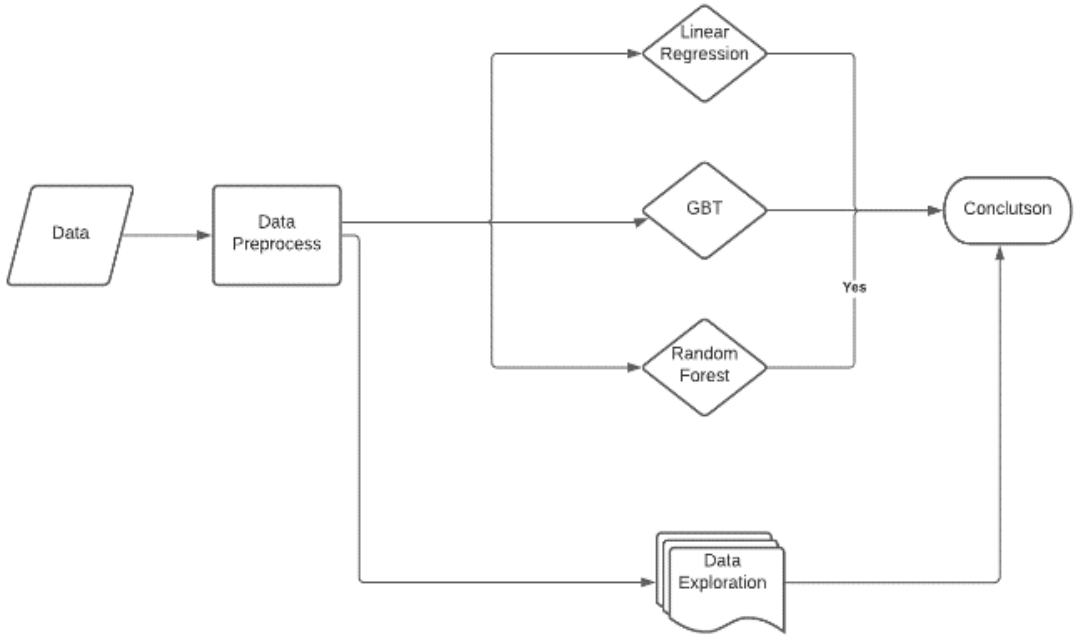
IST-718: NYC PROPERTY SALES DATA ANALYSIS

Abstract

- The coronavirus outbreak might have people feeling uncertain about whether a change in home prices will impact their housing plans. For property sales companies, what they care most is what kind of property may have the highest preference from the customers. Hence, we use the NYC Property Sales dataset to help predict the property prices.
- From the predictions, some inference questions can be answered, and they will be useful as reference for property sale companies.
- Several models will be applied to the dataset to find the influential factors on housing prices. Linear regression, random forest and gradient boosted trees are applied for our project purpose.
- After the data analysis, the main goal of the project is to find ways to calculate property sale price with property features as a reference for buyers and sellers when they are trading property. Companies will be able to make critical decisions related to hiring and investments based on the forecasting.

Methodology

- Our process follows the flow diagram. Data will be cleaned first. NAs will be removed and unreasonable data types will be regulated into the form qualified for the models inputs. The dataset will be separated into training and testing part. After that, three models, linear regression, random forest and GBT, will be applied to solve the regression problem. In the evaluation step, mean square error is utilized to grade the prediction accuracy of our models. With the help of the models and evaluation, we can draw our conclusion for this project.



Conclusion

- Our main goal of predicting property sales price is to provide a tool for people who want to buy or sell property. Using this tool, they can get a reasonable price if they input the property information. The random forest performs best.
- The most important feature of the property is the gross square feet, residential units, commercial units and building age. These are the factors people most care about when buying or selling property.
- The overall trend of the property sales price is ascending. The property of Manhantton and Madison has the highest price.

Syracuse University

Conclusion

Conclusion

- This portfolio has demonstrated the successful implementation of these learning objectives and the major practice areas in data science. Data was collected and managed using web scraping and application programming interfaces in conjunction with database solutions to be analyzed using statistical methods and data mining techniques for tasks such as regression, classification, or clustering. Various data visualizations were paired with clustering techniques to identify patterns which directed the respective analyses; actionable recommendations were developed to reflect tangible business decisions
- Communications skills were developed and displayed in the organization and delivery of insights, expressing them in terms which could be simply understood and acted upon. The ethical dimensions of data science practice were also reinforced in these applications by selecting only relevant data and considering user privacy when analyzing personally identifiable information. Syracuse University's School of Information Studies provides students the opportunity to synthesize the collection, management, and analysis of data, as well as the delivery of actionable insights using various data science techniques.

Syracuse University

Thank you