

portfolio milestone

Syracuse University, iSchool



Chensheng ma

SUID:243685555

[Company address]

1. **Introduction**
2. **IST 659 - Database Administration:**
   1. **Project Description**
   2. **Learning Goals & Reflection**
3. **IST 687 - Introduction to Data Science:**
   1. **Project Description**
   2. **Learning Goals & Reflection**
4. **IST 707 - Applied Machine Learning:**
   1. **Project Description**
   2. **Learning Goals & Reflection**
5. **IST 718 - Big Data Analytics:**
   1. **Project Description**
   2. **Learning Goals & Reflection**
6. **Conclusion**
7. **Personal Information**
8. **Introduction:**

The Applied Data Science program at Syracuse University’s School of Information Studies provides students a professional opportunity to collect, manage, analyze, and develop various business insights using data from a multitude of domains using various tools, language platforms, and techniques.

From curriculums of the program include Database Administration (IST 659,2021), Intro to Data Science (IST 687, 2021), Applied Machine Learning (IST 707, 2021), and Big Data Analytics (IST 718, 2021), these courses presented and reported different business insights by using SQL Server, Power, R, Excel, Tableau, and Python. These skills developed at the School of Information Studies furnish data scientists focused on 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.

1. **IST - 659 - Database Administration:**
   1. **Project Description:**

In this course project, we will develop a database to help Pet Hospital to organize names of doctors and nurses, patient recording, Date Time of the appointment, and patient kinds. Building conceptual and logical models to organize the relationships between the above variables.

Diagram

Description automatically generated Diagram

Description automatically generated

Tables were created in the SQL Server Management Studio while data population was accomplished using Microsoft Access. Then, the Microsoft Power application to help us to develop a beta version for the pet hospital.

* 1. **Learning Goals & Reflection:**

The progress of the project developed a solution that reveals the significance of ETL, that managers and data scientists can look through the data directly and efficiently. In the further course, we also learned how to build more complex and stable databases and how to visualize them in the business insights during the course Data warehouse.

This project contributed to a successful business application through collecting, organizing, and managing data. The hospital can directly check their patients' status and available appointments for their clients. As well, data scientists can use these data to build different statistical models for further investigation that finds business insights or care strategies for hospitals.

1. **IST 687 - Introduction to Data Science:**
   1. **Project Description:**

Through studying Data Analytics under the direction of professors Stephan Wallace and Jasmina Techeva, various data mining techniques were performed in data collection, processing, transformation, management, and analysis with different data science technologies, such as the SVM model, Linear Regression, and Data Visualization. In the final presentation, the SVM model and Linear Regression were implemented to compare visitor costs in the hotel by R Studio.

The results of this analysis were broken into segments based on the findings the dataset provided. The segments which are four different analytics, contributed to our approach to the data & gave answers to questions. The four different analytics are Descriptive, Diagnostic, Predictive, and Prescriptive. The report will reflect our thoughts for each individual approach.

People stay in hotels for a variety of reasons. Some stay because they are traveling for business, while others stay because they are on vacation. Every traveler has a certain amount of money they can afford to spend, each time they visit your hotel. Guests can spend money on a room, on food and drinks, as well as amenities and services. The total amount of money the hotel receives for all the goods and services it provides to its guests is defined as the Hotel’s Revenue

* 1. **Learning Goals & Reflection:**

Based on different data science technologies, we developed various recommendations for the hotel which can increase their business profit and visitor attendance. It provided an opportunity to practice how to translate data into business insights and support understanding to those customers who do not have data science and mathematics knowledge.

Recommendation 1: *Dynamic Pricing Based on Seasons*

The datasets provided a range of dates that travelers visited our clients' hotels. Creating dynamic pricing based on seasons would generate greater revenue based on peak traveling season. Dynamic pricing signifies a step toward a more efficient revenue strategy.

Recommendation 2: *Package Discounts*

The analysis has shown that there is a considerable number of customers who are frequent visitors combined with less likely to cancel their reservations. Providing those customers with package discounts would contribute to more stays in the future.

Recommendation 3: *Global Marketing*

Both locations have demonstrated customer exposure around the world. The country that has more visitors is coming from international locations. By concentrating on these countries, we can solidify our relationship with our visitors as well as expand our influence on neighboring countries.

Recommendation 4: *Overbook*

With a large number of cancellations, it will be smart to overbook to an extent to help ease the value gap for those customers with less lead team between cancellations. This approach gives each hotel the advantage of anticipating cancellations while also extending our marketing foothold.

Recommendation 5: *Staff Training*

The analysis shows agent consideration matters. Many of the canceled reservations were linked to several unique agent id numbers. This reinforces the notion that each employee has the power to make an impact, either negatively or positively. Preparing additional training for travel agents will work to increase additionally confirmed reservations as well as prevent more future cancellations.

1. **IST 707 - Applied Machine Learning:**
   1. **Project Description:**

In this project, we will be focusing on building machine learning algorithms to find and classify legendary pokémon. First, we need to know that most Pokémon games are battle-based. Although some people might be fond of collecting pokemon that look cute, most people would still like to get pokémon with very combat capabilities. There are some pokémon in Pokémon games that are legendary grade. These pokémon usually have high stats and are hard to catch. Hence, many people are interested in collecting these legendary pokémon since most of them have overwhelmingly strong combat powers.

* 1. **Learning Goals & Reflection:**

There are countless examples of people doing machine learning on animals and plants. People can classify animals and plants according to their traits and use their traits to predict what they are. What if we try something virtual using virtual monsters like pokémon? Pokemon are virtual animals or monsters that have their traits. Perhaps we can let machines learn to predict and classify different pokemon.

In this investigation, more than one thousand unique pokemon from 8 different generations only contain a little more than one hundred legendaries. Our target is to find those legendary pokémon; thus, we need to discover the correlation between legendary pokémon and other pokemons’ attributes. After creating some bar plots and histograms, we found that legendary pokémon have correlations with their stats, generations, catch rates, etc. Amongst all the correlations, we found that pokemons’ based stats have the strongest correlation with whether the pokemon is legendary or not. Then, the decision tree, clusterings, SVM, KNN, and the Random Forest models were used for this investigation based on pokemons’ stats. The decision tree roughly tells us the k-value can be 3 to 5. In the clusterings, we put pokémon into two and three groups, and legendary pokémon tend to belong to the groups that have higher overall stats, so the clusterings help to figure out legendaries have better stats than the normal pokemon. After using three-fold and five-fold cross-validation and repeated 10 times, the random forest model had the best accuracy for this investigation. We may consider using the random forest model for predicting the legendaries.

1. **IST 718 - Big Data Analytics:**
   1. Project Description:

In this project, we would like to apply our skills to make the trading firm Optiver better. Accurately predicting its volatility is essential for the trading of options, whose price is directly related to the volatility of the underlying product.

Problems we may have:

How to process outliers?

The real stock price is not in a stable area, it has a very crazy fluctuation(e.g., -10% - 10%). It means that we cannot just simply put historical stock prices into our model to do prediction.

How can we compare the price of a stock between yesterday and today?

How to get a good sample from the original dataset?

Since It’s a 2.5GB size dataset, we need to extract a small dataset as a sample.

* 1. Learning Goals & Reflection:

In financial markets, volatility captures the amount of fluctuation in prices. High volatility is associated to periods of market turbulence and to large price swings, while low volatility describes more calm and quiet markets. For trading firms like Optiver, accurately predicting volatility is essential for the trading of options, whose price is directly related to the volatility of the underlying product.

Building models that predict volatility and continuously generate fairer options prices for end investors.

Through this investigation, basic linear regression is not fit for our analysis. In the further goal, we need to consider adding more mathematical features based on the original dataset. Using a regression model with Lasso to find feature importance, we can develop a more accurate model based on these features. Due to the time series of our stock, we can use the RNN model and even LSTM model to reduce the score of realized volatility. This data analysis also leads us to learn how to use TensorFlow to run our models, since this competition dataset is large.

1. **Conclusion:**

This portfolio has demonstrated the successful implementation of learning projects and course practice in data science area. Data was collected and managed using application programming interfaces in conjunction with database solutions to be analyzed by using statistical methods and data mining techniques for tasks such as linear regression, classification, clustering, RNN, LSTM, and etc. And, how to use the EDA to investigate the dataset to discover patterns, and outliers, and form hypothesis based on our understanding of the dataset. A high quality of visualization will give clients and other data scientists a direct and plicit data understanding. This portfolio also showed the process of how to transfer the statistical report to the accurate business insights.

Through the learning process, programming languages, statistics, machine learning algorithms are just the foundation of how to become a data scientist with a master's degree. During the process, I got more valuable experience about corporation skills with different teams, analytical thinking, and accurate description of reports. My master program also fostered a data management practice with transparency, reproducibility, and ethical which promotes integrity within an organization’s analytics team.

1. Personal Information:
   1. Person Name: Chensheng Ma
   2. SUID: 243685555
   3. Email: [viscountvincent@hotmail.com](mailto:viscountvincent@hotmail.com)
   4. Degree: Master program in Applied Data Science at Syracuse University
   5. LinkedIn URL: [linkedin.com/in/vincent-ma-935366198](https://www.linkedin.com/in/vincent-ma-935366198?lipi=urn%3Ali%3Apage%3Ad_flagship3_profile_view_base_contact_details%3BQ7roqwlxTUaf9nOXK3oOCg%3D%3D)
   6. GitHub: <https://github.com/viscountma/M.S.ADS_Portfolio_Milestone.git>