Project Portfolio Milestone

College: School of Information Studies

Major: Applied Data Science

Name: Zeyang Zhou

SUID: 205387290

Email: zezhou@syr.edu

Github: https://github.com/wozhouwozhou/ADS-Project-Portfolio-Milestone

Table of Contents

1. Introduction

Applied Data Science, ADS is a Master program of School of information studies at Syracuse University. According to the description, The Master's in Applied Data Science is an interdisciplinary program that allows students to study a wide range of topics in data science.

Successful students in the Master of Applied Data Science program will be able to:

1. Describe a broad overview of the major practice areas in this 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 managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.

7. Synthesize the ethical dimensions of data science practice (e.g., privacy).

I'm on the old track with 36 credits and want to be a data engineer or analyst. During my two years of study, I learned a variety of talents and gained extensive expertise in data analysis, data science, data visualization, and other areas. In the portfolio paper, I will highlight tools that are really valuable to me, such as Mysql, RStudio, Jupyter notebook, MongoDB, Excel, Tablaeu, and programming languages like Python, R, and SQL. Furthermore, there are numerous useful machine learning and deep learning algorithms that allowed me to efficiently process so many data mining problems and got a high score data analysis competition from Kaggle. Because I completed 12 courses over four semesters, I decided to only display final projects from the following courses: IST 687 – Introduction to Data Science, IST 659 – Data Administration and Database Management, IST 652 – Scripting for Data Analysis, and IST 718 – Big Data Analytics.

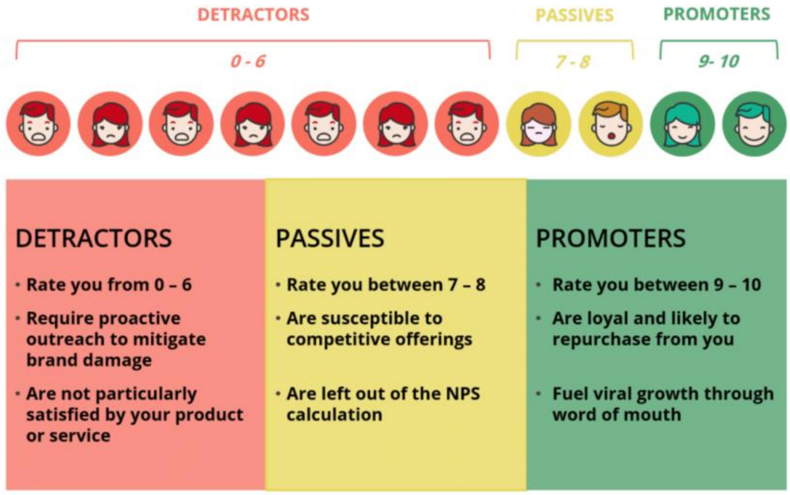
2. IST 687 – Introduction to Data Science

GitHub: https://github.com/wozhouwozhou/ADS-Project-Portfolio-Milestone/tree/main/IST%20687

The final project of IST 687 was a data analysis project, “Customer Churn in the airline Industry”. This project was in teamwork collaboration, and I finished it in my first semester (2020 Fall). Our main goal was to identify key patterns within this data and derive the best recommendation possible for Southeast Airlines to improve their customer satisfaction. The airline survey dataset was provided by Prof. Jeff who was our instructor. This dataset contains 88,000 survey responses from travelers who rated their overall satisfaction on a scale from 1-10. It was originally stored in JSON file and without data cleanse. So, in this research, we had all data analysis steps, except data collection, such as data exploration, missing value processing, data visualization, modeling and etc. The programming tool and language we used is RStudio and R.

First, we thought it was important to ask ourselves several business questions to keep in mind when doing our analysis. These questions are like “How can we improve NPS score?”, “How does NPS score vary by geographic location?”, “Depending on the data, who is a better partner?” and on. With these questions, we could find the direction for our data exploration. Next, we finished data Acquisition, cleansing, transformation, and munging. We checked to see if there were any missing data fields throughout the data set. We did keep in mind that if a field were not applicable, it did not necessarily mean the data was blank, but because the value was in fact meant to be 0. For example, we found that there were four missing values in the Likelihood to Recommend column, so we eliminated those blank values. For data transformation, we created a NPS calculation formula to get the accurate NPS score and it’s the figure1 below.

NPS Score: 8.9%



DETRACTORS

- Rate you from 0 – 6
- Require proactive outreach to mitigate brand damage
- Are not particularly satisfied by your product or service

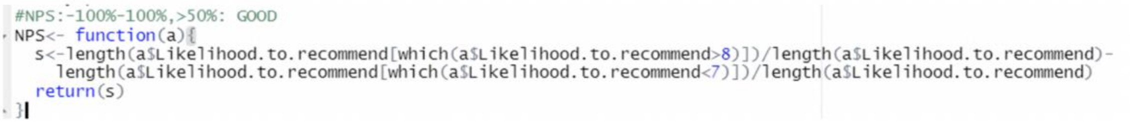
PASSIVES

- Rate you between 7 – 8
- Are susceptible to competitive offerings
- Are left out of the NPS calculation

PROMOTERS

- Rate you between 9 – 10
- Are loyal and likely to repurchase from you
- Fuel viral growth through word of mouth

NPS Function:



```
#NPS:-100%-100%,>50%: GOOD
NPS<- function(a){
  s<-length(a$Likelihood.to.recommend[which(a$Likelihood.to.recommend>8)])/length(a$Likelihood.to.recommend)-
    length(a$Likelihood.to.recommend[which(a$Likelihood.to.recommend<7)])/length(a$Likelihood.to.recommend)
  return(s)
}
```

Figure 1.

And then, for description statistics, we ran a basic summary of the data set to explore the descriptive statistics among the 32 variables to get an idea of what we were working with. We also focused these variables into extensive visualizations to get an idea of which areas we can deeply explore more. We explore the relationships between these variables and features like shown in the Figure 2.

Flight Delays

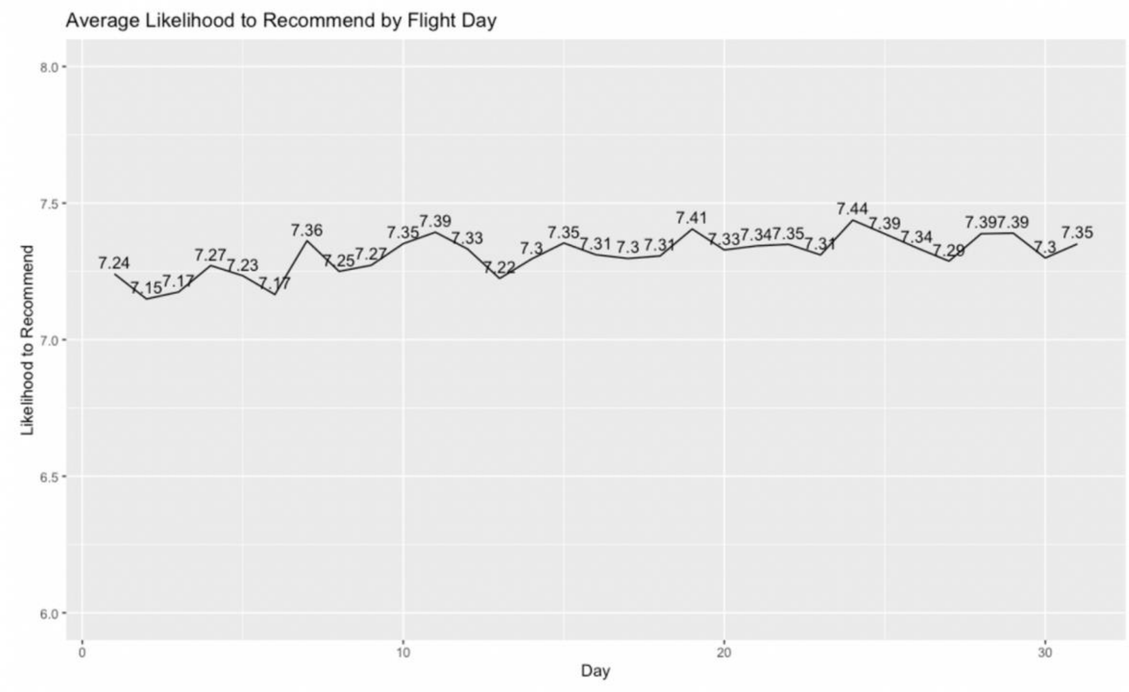


Figure 2.

This shows the average likelihood to recommend for specific days of the month. These are aggregate scores for the day number in January, February, and March (the months in this data set). Over the days in the month, the minimum likelihood to recommend score is 7.15 and the maximum is 7.44. Overall, there is very little variation across all days. To this end, targeting specific days may not be a priority area for reducing customer churn due to the lack of variance. Also, we tried to use some machine learning model to deeply explore the relationships between theses attributes and NPS score. These machine learning models were associate rule mining, linear model, support vector machine and text mining (Figure 3). We really got something like “Age” has strongest impact on NPS score. But unfortunately, since I and my teammates were all first-year grad student, we failed in SVM and got just little from linear model. At last, we provided five t helpful recommendations for Airline SouthEast.

Text Mining

Finding out if the feedback was negative or positive

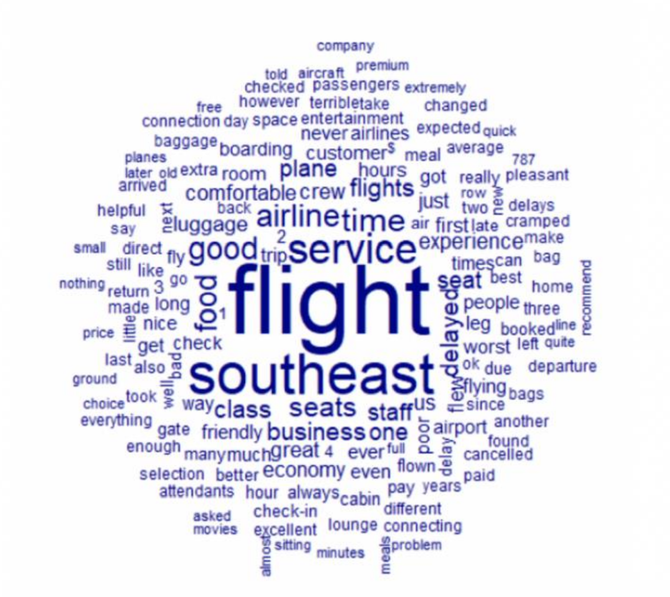


Figure 3

In a nutshell, this was my first research assignment involving data analysis at Syracuse University. To be honest, I had no prior experience in data science before to this assignment. I had no idea what R, RStudio, or data analysis were. This course, IST 687, provided an excellent foundation for me, allowing me to gain a wealth of information in the subject of data analysis, including but not limited to statistics, fundamental math, visualization abilities, machine learning, and how to build data models. Most of these complex challenges in progress are easily solved thanks to Prof. Jeff's thorough instructions. It is, in my opinion, not simply the principles or knowledge in the textbook, but the real-world data analysis practice program. I needed to interact with all of my teammates to stay up to date on everyone's progress, learn how to debug in RStudio, and schedule appropriate meetings because we were in four separate time zones (China, Korea, the United States, and India) because to the Covid-19 pandemic. Finally, I'd like to thank all my teammates and professors once more.

3. IST 659 – Data Administration and Database Management

GitHub: https://github.com/wozhouwozhou/ADS-Project-Portfolio-Milestone/tree/main/IST%20659

IST 659 covers relational database administration concepts and how to query using SQL. The final project was a group project, and I collaborated with one classmate to accomplish our project, the Covid-19 information management system. Unlike in IST 687, Prof. Fudge did not assign us a project to complete; instead, we were just required to use Azure Data Studio and Power App (Microsoft) to complete our backend and frontend material. It is a project that includes database design, business analysis, Entity-Relationship requirements, conceptual and logical model, and UI design in addition to SQL.

According to previous emails from SU, we can deduce that gathering and clustering are responsible for a significant increase in COVID-19 instances. As a result, we'd like to use this program as much as possible to limit the chance of infection. At the same time, it can keep the university open as usual. First, we created an ER diagram to find clear relationships between entities (Student, Test, Report, Application, Fever Record and etc.,) with their attributes such as Student – Test and Student- Report. Next, we constructed the conceptual model and logical model to deeply explain their relationships and functions which can be helpful in database design, as shown in the Figure 3 and Figure 4. Third, we completed the database and SQL query based on our user story including student and admin. Take “Student” as the example, a student can submit their destination, date of their travel so that an admin can judge whether the student is from a high-risk area and determine to change his/her status to red or not. (Duration > 10, status: green=>red). Or A student should submit a request so that he/she can participate in an activity more than 5 people (First, his/her request will be checked automatically by a trigger. If his/her student status is red, his/her request will be refused automatically. And then, admins check the other requests). Finally, we connect the UI interface to the prepared database, which contains ten tables.

Ultimately, IST 659 is an excellent course for my database foundation. I learned a lot about database administration fundamentals and SQL syntax. Database management is essential for a data engineer in the field of data analysis. It is due to the electrical business style, which requires managing hundreds or thousands of customers' information. Nowadays, there are many different types of databases for video, picture, and audio, including as MySQL, mongo dB, and certain NOSQL databases. With the assistance of this project, I can go deeper into these databases in my second and fourth semesters.

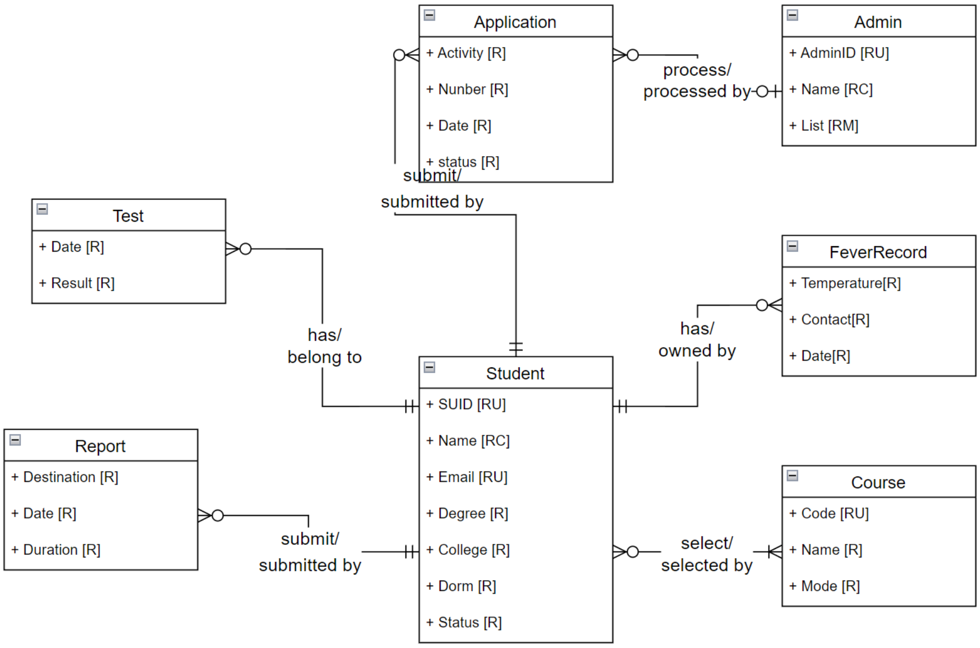


Figure 3. Conceptual Model

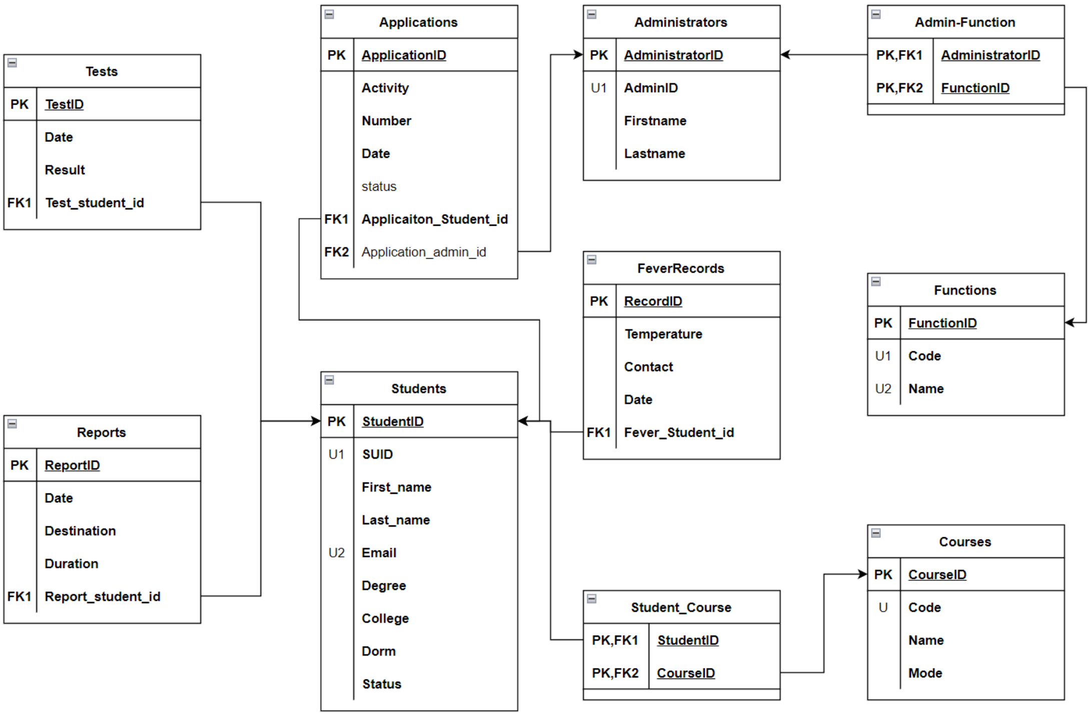


Figure 4. Logical Model

4. IST 652 – Scripting for Data Analysis

GitHub: https://github.com/wozhouwozhou/ADS-Project-Portfolio-Milestone/tree/main/IST%20652

IST 652 was a course about how to use python to do data analysis. It's in group work and I had four team members. Because this course was based on python, we were required to do our data analysis in Jupiter notebook with several python packages taught in class like pandas, NumPy, matplotlib, sk-learn and etc. According to project description, "The project focuses on open data in order to ensure that your chain of transformations and analysis is reproducible, we chose a hotel booking demand dataset from Kaggle as our dataset. The total number of online travel bookings made each year is around 148.3 million, which generates sales of around $755 billion per year in 2020. And it continues to increase since 2014, with an average of 10% every year. But in the real world, many people are troubled by how to book a value hotel. It's no surprise that over half the people spend more than one week researching their hotel before the holiday. That's why we chose this topic.

Initially, we explore our dataset and do data cleanse. This data set contains booking information for a city hotel and a resort hotel and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. we processed missing values in this dataset. There are 4 columns with missing values which are “agent”, “country”, “children” and “company”. Since “company” has too many missing values and it is less helpful to our analysis, we decide to delete all of them from the dataset. In column “children”, we fill the missing values with the mean value of “children” since it’s just 0.0035% missing values. In column ‘country’, we fill the missing values with the mode value. In column ‘agent’, we replace the missing values with 0 because these customers may select independent travel and they do not need to report the IDs of their agent. We also transformed features with the incorrect datatype to their corresponding ones. Then we dealt with outliers, duplicated values, and replaced values. Following that, we visualize our data using a variety of plots such as histograms, bar plots, boxplots, and scatterplots. In Figure 5, a scatterplot is depicted.

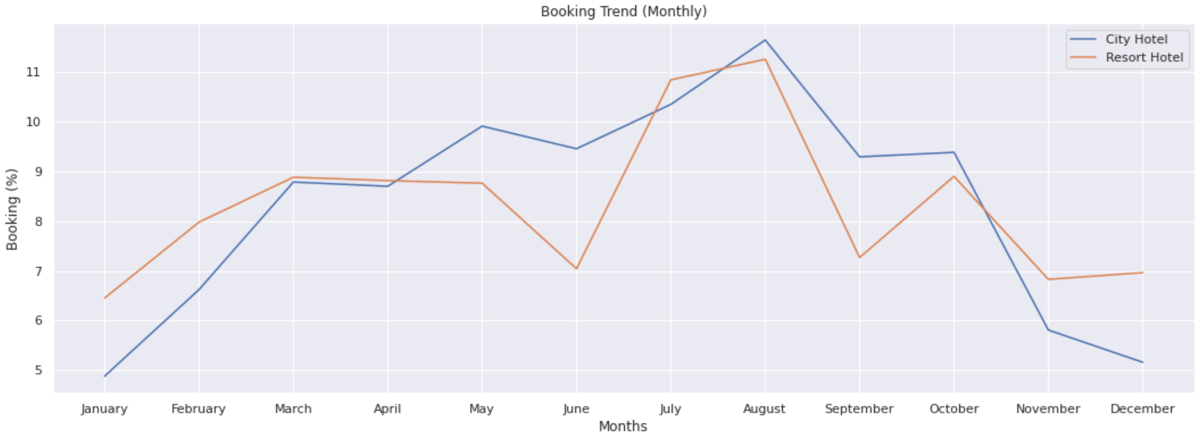


Figure 5.

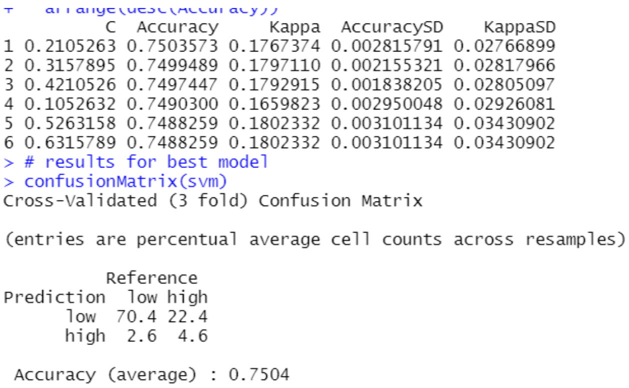
From the plot above, we can see that the busiest months are July and August. I believe the reason could be there are beautiful sceneries and stunning sunshine in summer, when people love to hang out with families and friends. The least bookings were made from January, November and December, which are the coldest months. Finally, we do business analysis for the hotel booking market using the plots and models we developed. To determine the relevance of a feature, we utilize a logistic model and a random forest model. Given that our objective variable is either a resort hotel or a city hotel, we believe logistic regression is one of the better models to use. We used the accuracy score to evaluate the performance, and the logistic regression model had an accuracy of 80.25 percent. We also plot the ROC curve for logistic regression, and the AUC is 0.88, which is excellent. We came to a few interesting findings. For instance, resort hotels tend to have less bookings than city hotels, therefore they need to improve their marketing approach and advertise the hotels more, especially on social media. I learned how to do data analysis using python, the most useful data science tool, as a result of this training. What was also crucial in the lesson was the use of reptile in python. I can send a request to any website containing data of interest and obtain the entire dataset in the file format of my choice.

5. IST 707 – Data Analytics

The final project of IST 707 is a data mining project based on R and the tool required by Prof. Yang was RStudio. This is my first-time approaching data mining skills. It was my personal project. The goal of this project, as IST 652, is to use the primary abilities learned in this semester to solve a real data mining problem and choose my own dataset. I learned a lot about regression and classification methods such as KNN, K-means, and SVM with various kernels. Machine learning is also an important aspect of the coursework. I fully understand what data mining is and when and where we need to utilize whatever model or algorithms because of studying these.

The dataset I chose is from Kaggle (2016) which contains a list of video games with sales greater than 100,000 copies in different areas from 1980 to 2016. It was generated by a scrape of vgchartz.com. The trend of the desire to take video games as a leisure time entertainment has been intensified in recent times. Due to Covid-19, video game market is becoming more popular than before since it changes the lifecycle of most people. Now the gaming industry is booming and, according to Global Industry Analysts, the global market for Video Games estimated at US$156.8 Billion in the year 2020, is projected to reach a revised size of US$293.2 Billion by 2027, growing at a CAGR of 9.3% over the analysis period 2020-2027. This final report will show a complete procedures of data mining with machine learning models including data preprocessing, descriptive analysis, visualization, and analysis for results generated by different machine learning models. Four machine learning algorithms planned to use are KNN, SVM, Naïve Bayes and Apriori.

First, I looked through the dataset. The dataset in this example has 16598 records with 11 variables. These works allow me to detect outliers, process missing values, convert datatypes, and preprocess them for machine learning. Second, I split the raw dataset into two subsets based on the timeline because I would like to train the machine learning model to predict hit video games in the future. Second, I divided the raw dataset into two groups based on the timeline since I want to train the machine learning model to predict future hit video games. One is the training set, which contains 60% of the data, and the other is the test set, which contains the remaining 40% of the data. It's time for me to build the four machine learning models and obtain the ROC curve and prediction accuracy. Take SVM as an example, the ROC curve and accuracy of it is show in Figure 6 and Figure 7.



```
+    arrange(desc(Accuracy))
          C  Accuracy     Kappa  AccuracySD    KappaSD
1 0.2105263 0.7503573 0.1767374 0.002815791 0.02766899
2 0.3157895 0.7499489 0.1797110 0.002155321 0.02817966
3 0.4210526 0.7497447 0.1792915 0.001838205 0.02805097
4 0.1052632 0.7490300 0.1659823 0.002950048 0.02926081
5 0.5263158 0.7488259 0.1802332 0.003101134 0.03430902
6 0.6315789 0.7488259 0.1802332 0.003101134 0.03430902
> # results for best model
> confusionMatrix(svm)
Cross-Validated (3 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction  low high
      low  70.4 22.4
      high  2.6  4.6

 Accuracy (average) : 0.7504
```

Figure 6.

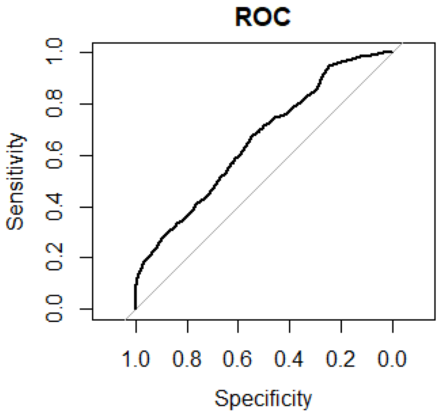


Figure 7.

Finally, based on the table, SVM with RBF kernel is proved to be the best classification machine learning algorithm for this study. However, it's ROC sensitivity is just 0.5454 which means the result of it is not stable and almost random. Similarly, ROC sensitivity of SVM with different kernels all do not perform well which may be caused by the dataset. To small dataset, KNN may be a better choice.

To sum up, because this is my first time completing a complete data analysis report based on machine learning, many issues have arisen, such as how to choose a topic, errors in R Studio, analytical logic, and so on. The great power of machine learning impressed me the most in this endeavor. To solve various data analysis challenges, I can apply various machine learning techniques. However, my lack of machine learning experience prevents me from performing well in this final assignment. So, in my future learning, I plan to pursue further research on how to use machine learning models to solve real-world business challenges or to assist in business consulting.

References

Applied Data Science Master's Degree. (n.d.). *iSchool | Syracuse University*. Retrieved February 18, 2022, from https://ischool.syr.edu/academics/applied-data-science-masters-degree/

Program: Applied Data Science, MS - Syracuse University—Acalog ACMS™. (n.d.). Retrieved February 18, 2022, from http://coursecatalog.syr.edu/preview_program.php?catoid=18&poid=9483&returnto=2322

ADS-Project-Portfolio-Milestone/README.md at main · wozhouwozhou/ADS-Project-Portfolio-Milestone. (n.d.). GitHub. Retrieved February 18, 2022, from https://github.com/wozhouwozhou/ADS-Project-Portfolio-Milestone

Hotel booking demand. (n.d.). Retrieved February 19, 2022, from https://kaggle.com/jessemostipak/hotel-booking-demand

wozhouwozhou. (2022). ADS-Project-Portfolio-Milestone [Jupyter Notebook]. https://github.com/wozhouwozhou/ADS-Project-Portfolio-Milestone