

Name: Chenghao Wu

SUID: 222142826

Email: [cwu137@syr.edu](mailto:cwu137@syr.edu)

Portfolio:<https://drive.google.com/drive/folders/1N0Frpn4AIJj00D1YVS9RsdjK-vrozV4W?usp=sharing>

Applied Data Science

Portfolio Draft

**Contents**

[**Introduction** 2](#_Toc53999569)

[**Projects** 2](#_Toc53999570)

[**IST 718 – Pricing analysis and prediction of Airbnb in Seattle** 2](#_Toc53999571)

[Data Collection: 2](#_Toc53999572)

[Data Cleaning 2](#_Toc53999573)

[Data Analysis 2](#_Toc53999574)

[Strategy and Decision 3](#_Toc53999575)

[Ethical Dimension: 3](#_Toc53999576)

[**IST 707 - Analysis of Default Payment in Credit Card** 3](#_Toc53999577)

[Data Collection: 3](#_Toc53999578)

[Data Cleaning: 3](#_Toc53999579)

[Data Analysis: 4](#_Toc53999580)

[Strategy and Decision 4](#_Toc53999581)

[Ethical Dimension: 4](#_Toc53999582)

[**IST 664 – Hotel reviews rating prediction** 4](#_Toc53999583)

[Data Collection: 4](#_Toc53999584)

[Data Cleaning: 5](#_Toc53999585)

[Data Analysis: 5](#_Toc53999586)

[Strategy and Decision: 5](#_Toc53999587)

[Ethical Dimension: 5](#_Toc53999588)

[**Conclusion** 5](#_Toc53999589)

# **Introduction**

This protofolio’s objective is to demonstrate the learning outcome of the Applied Data Sciencie program, including demonstrations of the ablilty of describing a board overview of the major areas in data science,collecting and organizing data, identifying patterns in data via visualiztions, statistical analysis as well as data mining. There are three projects will be demonstrated in this portfolio. In each project , multiple business decisions and stratigies will be concluded based on the result of analysis. What’s more, I will discuss the the ethical dimension of the following projects to better meet the goal of data science study.

# **Projects**

## **IST 718 – Pricing analysis and prediction of Airbnb in Seattle**

Airbnb has become a worldwide application that people use broadly during personal or business travels for its convenience and cost-effectiveness. This project is to identify the dominant factors of the house price in Airbnb. It also analyzed the sentiment of guests and provided market strategies to increase the revenue for the stakeholder of Airbnb. To better apply our project in real industrial, our team designed a simplified recommendation system for Airbnb users as well.

### Data Collection:

The Airbnb datasets are available at[**http://insideairbnb.com/seattle/**](http://insideairbnb.com/seattle/). This website is an independent, non-commercial set of tools and data that provide a lot of the data of Airbnb which is really being used in cities around the world. There are three individual dataset we are going to analysis in various perspectives. Among them, the listing dataset has a comparatively large number of attributes,which contains 106 columns and 9024 rows ,so we categorized the columns based on their property at the very first step. Reviews dataset has more than 40,000 records about the reviews of houses from 2008 to 2019, calendar indicates the availability of a particular house, which would not be used in this project

### Data Cleaning

Since we had a massive dataset in the project, the process of data cleaning is a good way to demonstrate how should we deal with data that closed to our real life. We used multiple data visualization techniques such as barplot, word cloud and boxplot to explore the dataset itself. Then we drop some columns having half of missing reocords since we don’t want introduce any bias. The NA’s in Boolean columns are replaced with mode, and the mean for numerical continuous features.

### Data Analysis

Based on business understanding, we selected the features that have a strong correlation to the housing price. By streaming the analyzing process, the filtered dataset will be fed into the data pipeline. For a particular model, we reduce the dimensions of the dataset, impute the missing value and outliers, scale and visualize the distribution of each feature to check the effectiveness of the data cleaning process. For descriptive analysis, we used data visualization methods to find the pattern of the dataset and get a insight of the business question we are trying to resolve. For predictive analysis, we used Lasso regression and tree-based algorithm(random forest and GBT) to find the most important features that have big impact on the price. We also used logistic regression to perform a sentiment analysis based on users reviews. Finally, we used clustering algorithm (K-means) to build a recommendation system via using TFIDF.

### Strategy and Decision

Based on our analysis, we proposed several business advices for both Airbnb stakeholders and users. It is very important that to interpret the result of data manipulation. A skill of communicating with relevant professionals cannot be emphasis too much. In this project, we advice the stakeholder to improve the living environment and the class of Airbnb. Meanwhile, a recommendation system is also useful for Airbnb users and have a big business value.

### Ethical Dimension:

The dataset we used contains a lot of personal information of Airbnb hosts, but those are public information to better introduce their house. For Airbnb users, there are few information of them besides their reviews since the Airbnb has to protect user’s pravicy. We can only obtain their reviews left for hosts but not the personal data such as phone number or their age.

## **IST 707 - Analysis of Default Payment in Credit Card**

The default payment analysis is a task that for bank to predict potential risk of their client’s default payment by finding out the important factor. In this project ,we used three different models , which are SVM, random forest and neural network, to predict if a client will have default behavior not. Besides, we also tune the models and evaluated each model’s performance and finally create a shiny app interface to let people do the same analysis without coding by themselves.

### Data Collection:

We collected the dataset from Kaggle website. This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

### Data Cleaning:

Before we did the modeling part, we took a glance at the dataset and found out some of the data need to be cleaned to fit our model. At the very fisrt, we removed all irrelevant columns since they will contribute little to the prediction. Then we converted category values to dummy variable. A feature correlation analysis was also performed to find out the rudent attributes, we use PCA to deduct the dimensionality to better impute the data. Finally, we did the data standardization and normializtion for some features.

### Data Analysis:

Our team member did data exploration by using ggplot2 to plot multiple graphs to demonstrate the overall information of the bank clients, including the composition of their gender, age as well education.Then we balanced our data since the proportion of no default records occupied 80% of whole data. After feature engineering, we trained the data in SVM , random forest and Neural network model and find the best hyperparameter for each model by fine tune them. We found out the most important feature by using random forest algorithm and compared the AUC of each model. Eventually we build a shiny app to demonstrate our analysis outcome.

### Strategy and Decision

Clients’ behavior prediction could be a most valuable part for any company. By analyzing clients’ previous actitivity, we can control our risk and make a optimal strategy. In this case, we found out that 2nd payment is the most important feature to predict if a people are going to default or not. We can suggest credit card bank to pay attention to their clients’ early payment credit.

### Ethical Dimension:

The topic we chose for this project is comparatively sensitive, however, the data itself doesn’t include any personal information of customers. We just tried to find out the pattern of customers’ behavior and use the outcome to protect company’s interest. People who has default payments will not only hazard their own credit, but also does harm to the society, so a research in terms of clients behavior is necessary.

## **IST 664 – Hotel reviews rating prediction**

Hotel reviews is a bridge between hotel managers and customers, from where the hosts can get feedback from guests to improve their service. The goal of this project is to using natural la guage processing techniques to dig valuable information from text data. Natrual language processing is a field of data science with huge potiential. By acquiring the ability to analysis text data, we can contribute a lot to our daily life. In this project, we will build several models to predict rating score based on reviews, and compared accuracy and F1 score by using unigram ,bigram, unigram and bigram as features in different classification models.

### Data Collection:

The dataset is about hotel reviews from different states in the United States and is provided by [Datafiniti Business Database](https://datafiniti.co/products/business-data/). It includes 2000 hotels and 10000 reviews updated between January 2018 and September 2018. This database contains 25 columns. (Hotel ID, Date Added, Date Updated, Address, Categories, Primary Categories, City, Country, Keys, Latitude, Longitude, Name, PostalCode, Province, Review.date,Reviews Rating, Reviews. User City, Reviews.User Province, Reviews.username, Source URLS, Websites). The dataset source is from Kaggle:<https://www.kaggle.com/datafiniti/hotel-reviews#Datafiniti_Hotel_Reviews.csv>

### Data Cleaning:

To impute the raw dataset, we firstly remove the irrelevant columns and only keep the columns of review and its rating star. Besides, we omit all NA’s in our dataset since we have a huge tokens in each review so the sample would be enough to analysis. We also found that some rating score are not rounded, so we convert them to integer data type and round them.Then we tokenized the review text, removed the stopwords, non-alpha tokens, and lowercased the tokens for further imputation.

### Data Analysis:

This project is mainly a multiple classification task since we have five classes of rating score, so we chose to use Naïve bayes, random forest and Linear SVC model to achieve our analysis objective. After using barplot and word cloud to do the exploration part, we firstly use the bag-of-words technique to apply navie bayes on most common 2500 words to train our model and to predict , then evaluate the model by using F1 score. Then we noticed that the data is imbalanced ,so we chose to use oversampled technique to balance our data to get a more promising outcome. We used three different features, which are unigram, bigram and both unigram and bigram in random forest and Liner SVC model. We found the linear SVC model has the best performance among three.

### Strategy and Decision:

By analyzing the frequency of the words in hotel guests review in different rating class, we found out that people tend to give a higher rating score when they got a clean living environment ,had a good breakfast or the hotel location is convenient. Contrastly, low rating score tends to appear when the room is dirty.

### Ethical Dimension:

In hotel review dataset, there is no guests’s personal information is used in analyzing process. We only focused on the relationship between the review and rating score. However, the bias may occour if some guests leave a low rating score or frustrating review on purpose to hazard some hotel reputation. The reviews are more human-centered infromation, because of that, we need to pay attention to reliability of the data source to prevent a misleading interpretion .

# **Conclusion**

The portfolio demonstrated the ablilty that I obtained from the study in Applied Data Science program, including collecting and organizing data, identifying patterns in data via visualiztions, statistical analysis as well as data mining. I also learned how to interpret the result of my analysis and apply them to business industrial. By using descriptive/ predictive analysis technique and the knowledge of communicating skill, I believe that I am able to handle various data analytic tasks individually now.