Applied Data Science Portfolio Essay

Mark Gopani
SUID:726681243
mgopani@syr.edu

Portfolio Link: https://github.com/MarkGopani/Applied-Data-Science-Portfolio

Table of Contents

Introduction

In the upcoming world, data is the new valuable resource. All industries of various sizes generate data on a regular basis, be it a small food truck start up or a big bank. The collected data can be put to use to drive the business. Data becomes an important asset of any organization in the sense that studying it helps understand and enhance the underlying process thereby saving time and money. In other words, it has become the backbone of business decision making.

In recent years the demand for skilled professionals in the field of data science has grown remarkably. Data science has become one of the most prominent career choices for new graduates as well as working professionals in all domains and hierarchies.

My engagement with data started early in my life when I took up coding as an elective subject during secondary school. I worked with java applications which immediately helped me recognise my desire to pursue programming and tech. After learning a few coding languages and completing high school side by side, choosing Information Technology as my specialization for my undergraduate course was an easy decision.

I was acquainted with programming and using different software to work with data during my 4 year undergraduate program. One of my projects during the program helped me formulate a path towards data science. In the project I dealt with data collection and then working with the data to achieve the goal of my project. The data collection part was the most challenging part however that is what drove me to dive deeper into handling data. Collecting it from multiple sources and then manipulating it to find hidden patterns intrigued me. I had assessed myself as a logical thinker and it helped me choose my next career checkpoint at Syracuse University where I enrolled for the Master's in Applied Data Science program.

Over the two year program at Syracuse, I was able to work with real data sets and apply all my previous knowledge with practical scenarios. I learnt a lot about cleaning and collecting data, making predictions using historical data, and interpretation of different analyses that can be done on the data.

I enrolled for courses which helped me understand business and finance. I learned techniques which could be applied to scenarios to get a rapid evaluation and understanding of the status quo, identify problems and find solutions

through the data available. I learned Natural Language Processing which helped me deal with text data and work on sentiments of text data from different data sources.

In the portfolio I give a precise description of how I delved myself into data science with the help projects that I undertook during the program.

Learning outcomes

The master's in Applied Data Science is an interdisciplinary program provided by the Information School and the Whitman School of Management at Syracuse University

The following points outline the basis of the learning outcomes:

- Collect, store, and access data by identifying and leveraging applicable technologies
- Create actionable insight across a range of contexts using data and the full data science life cycle
- Apply visualization and predictive models to help generate actionable insight
- Use programming languages such as R and Python to support the generation of actionable insight
- Communicate insights gained via visualization and analytics to a broad range of audiences
- Apply ethics in the development, use and evaluation of data and predictive models

In the following sections, I elaborate the Projects that helped me achieve the learning outcomes and learn deeper about certain areas of my interest.

Applying ethics in the development, use and evaluation of data and predictive models

Financial Analytics is one of the subjects provided by the Whitman School of Management which really got me interested. Before taking the course, I did not have abundant knowledge about the stock market and its workings. During the course I was introduced to different models and tools that could be implemented on stocks data for decision making.
I learned to implement and test the Capital asset pricing model which is a core topic in the finance domain. I learned to manage my own Portfolio with the help of my final project, I carried out risk premium estimation and performance evaluation of my own portfolio. My investment strategy helped me rank #5 among my excellent class peers. I also wrote a corresponding report on the portfolio investment strategy which helped me demonstrate time series modelling skills with the help of data visualisation and presentation in R.

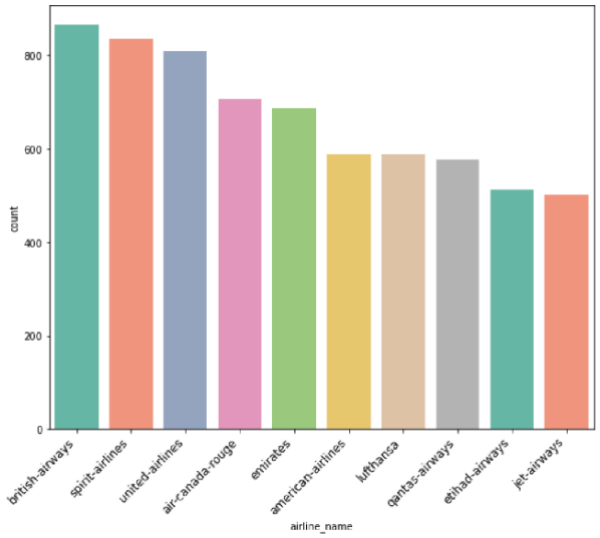
I enrolled for the summer course Natural Language Processing to increase my knowledge in the domain. I had previously come across a similar topic in my scripting class where we learned to identify text patterns and extract text corresponding with those patterns. It was interesting to find the how the topic could be adopted with respect to large amounts of text data that was available online in the form of product reviews or user comments and posts.
Natural Language Processing was taught to us by Professor Michael Larche. His comprehensive material helped me delve into topics like linguistic analysis, parts of speech tagging, sentiment analysis and word level semantics. Using the nltk package in python, I was able to master my command on text extraction, tokenisation while dealing with very unstructured data and filter out data which would be unethical in certain context and transform such data using python tools.

Airlines Review Analysis | IST-707 Applied Machine Learning | Python, Orange3

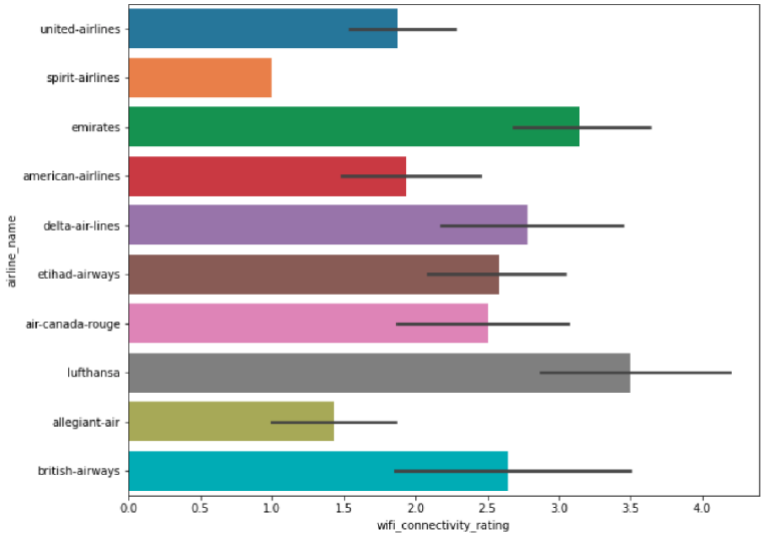
I was accustomed with various data mining techniques during the course of instruction. I familiarised myself with the Orange3 software environment, it was very simple yet efficient in conducting wide range of data exploration, visualization, pre-processing and modelling techniques. Writing python code in modules inside the model elements made modelling quick and efficient.
This project was undertaken by a team of two. Me and my classmate decided to explore the airline industry as a part of our final project under the guidance of Professor Joshua Introne. We went ahead with this topic because we were keen

about aviation, also the topic would help us implement our learnings from class as we had found an appropriate unstructured dataset which we believed would have hidden information to dig out.
The main focus of the final project was to implement data mining techniques to fetch and clean the data and then formulate business questions based on the exploratory analysis. Once we had determined our topic of interest after the exploratory analysis, we performed sentiment analysis and predictions.
In the first stage of the project, we cleansed the data by first removing null values and corrected the country name variable format. Our initial stage of sentiment analysis included cleaning the reviews column of the dataset.
In our next step, we performed exploratory analysis on the services provided by different airlines and also assign polarity to reviews given by customers.

We look at top 10 airlines which provide inflight entertainment:



Next we look at the top 10 airlines with best wi-fi connectivity ratings:



We then implemented supervised machine learning models to determine factors which would be most favourable for customers to give positive reviews for airlines and predict whether a certain airline was probable to be recommended or not by a customer.
With the help of this project I was able to apply my concepts of Natural Language processing on the reviews given by customers and also understand the associations of different variables with the help of regression and decision trees.

Managing data by identifying and leveraging applicable technologies and create actionable insight across a range of contexts

The course Database administration concepts and database management taught us the fundamental storage models and database concepts. We applied the database development life cycle in our project.
The project helped us learn writing complex queries to arrange data in standard storing norms and also fetch the data to perform qualitative analysis. We got exposure to Microsoft powerapps where we designed the frontend for a database app and integrated a menu to perform different data manipulation tasks such as fetching, filtering, sorting and validation with the help of queries.

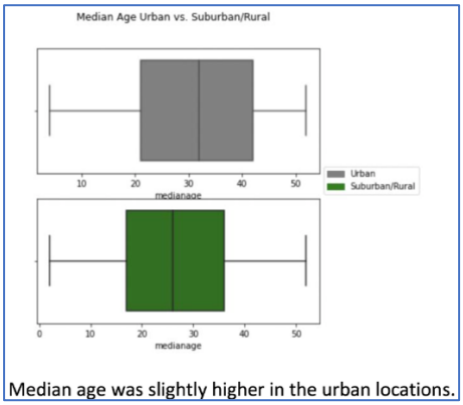
One of the core courses of the curriculum: Scripting for Data Analysis familiarized me with the concept of data pipelines, their importance to reduce redundancy and errors and how to implement them. In the course we focused on concepts like data collection where we performed data wrangling: learning to collect data with different formats.

We wrote scripts to access data of structured and semi-structured forms and define and find patterns in unstructured data. We carried out data validation and testing using python.
As a part of our course, we undertook two mini projects for which we framed real world business questions and attempted answering those questions with visualization analysis using python.

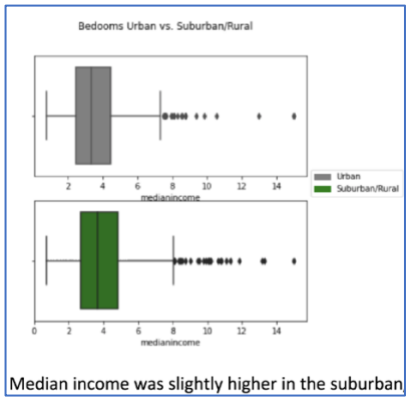
House Price Prediction | IST- 718 Big Data Analysis | Python

This project focused on analysing the real estate market of California. My team decided to explore this domain as we had seen similar examples in our class taken by Professor Willard Williamson. Our interest arose from the fact that we could fetch information from a dataset and try to predict the house price of a house which was not on the market using the variables that were recorded in the historical data.
This project aligned with one of the projects that I had done earlier with respect to making predictions using historical data. The real challenge for me was to carry out similar analyses but constraining myself to the coding language and the methodologies that were followed by our professor during the course of our class. The class helped us learn supervised and unsupervised machine learning

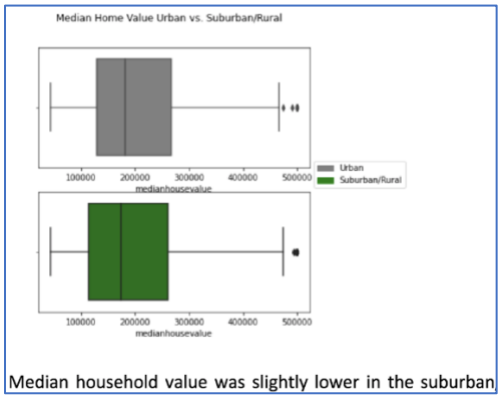
models along with in-depth model tuning and analysis. We carried out all our analyses in python using resilient distribution datasets exclusively.
We split the dataset into urban and suburban locations to compare house prices in the two areas considering the variables which we filtered out as most important with the help of principal component analysis.



Median age was slightly higher in the urban locations.



Median income was slightly higher in the suburban



Median household value was slightly lower in the suburban

The different models we trained for linear regression helped us understand the tuning parameters in the process and improve our overall results for the model.

Model	R^2	RMSE
Linear Regression	0.55	0.78
Linear Regression with parameter tuning	0.54	0.77
longitude, latitude, total rooms, and total bedrooms as input	0.30	0.98

Applying visualization and predictive models using programming languages such as R and Python to support the generation of actionable insight

The two courses that I took up under the Whitman school of Management: Marketing Analytics and Data analysis and decision making helped me get a better understating of applying visualisations to data analysis. It would be impracticable if we just perform analysis and then are not able to draw a clear picture of the undergoing processes. Marketing analytics helped me get familiar with different tools like R commander, XLSTAT, Ms Access.

Data analysis and decision making strengthened my command on R Studio, I learned to conduct analysis on populations and samples and draw inferences in the form of charts to get a visual of the analysis. I performed statistical tests including, t-test, z-score outlier evaluations, p-test, and hypothesis tests. At the end of the semester after taking these courses made me feel confident about my comprehension of conducting analysis and showcasing my findings to probable stakeholders with the help of visualisations.

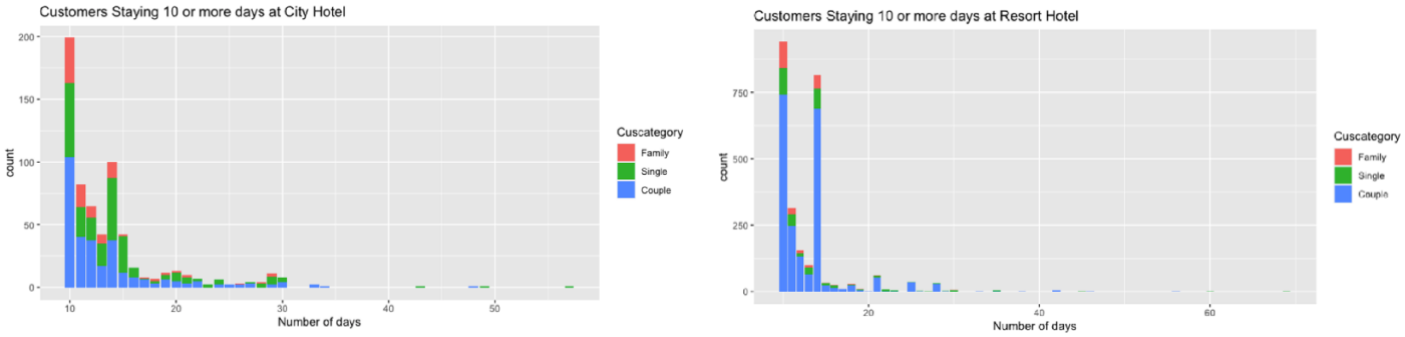
Analysis of Hotels Dataset | IST-687 Introduction to Data Science | R programming language

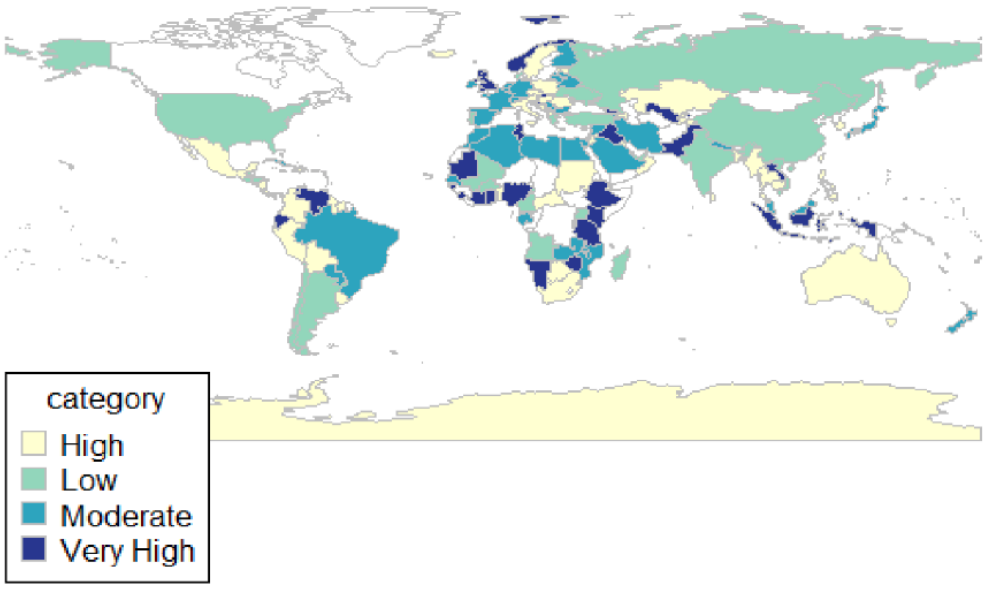
Professor Ayse Dalgali introduced us to the real power of performing machine learning in Rstudio: and open source software. We learned scripting in R studio for data management and visualization. We explored different data management techniques including linking, aggregation, summarisation and searching. We learned to determine appropriate techniques to analyse data and apply them at different stages of a project lifecycle.

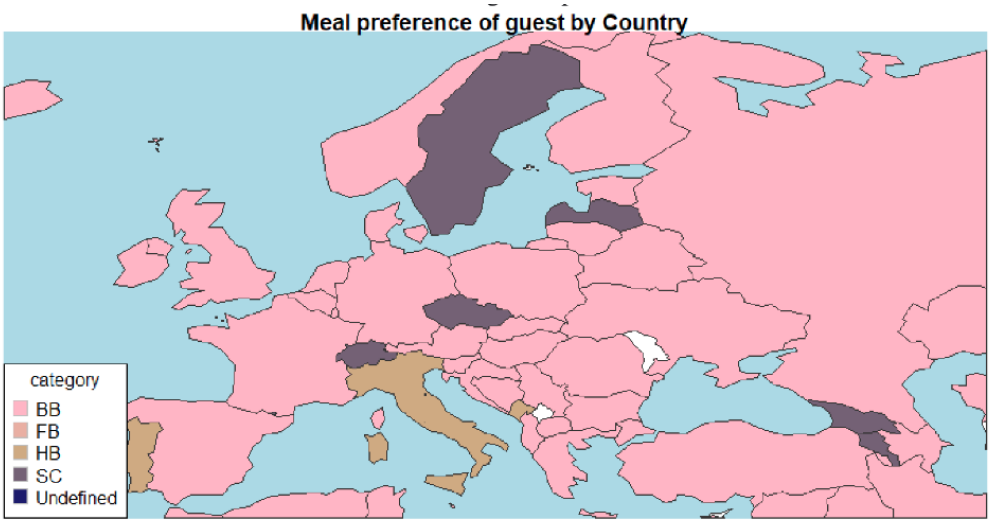
The main project focuses on exploring two data sets which constituted booking data of two hotel resorts at geographically different locations in Portugal. One hotel was amidst a hasty city in Lisbon and the other Resort in a secluded area in Algarve.

The team comprised of me and 2 other classmates. We explored the dataset initially to find any irregularities. We transformed all the data to a format which would make it easier for us to work on it for fitting it into models. The data had to be cleansed a lot before we could use it for the same: we carried out NA-interpolation and then introduced new column to view the data by season.

The analysis that we wanted to perform focused on customer segmentation and seasonality analysis for revenue generated and cancellations among the two hotels. We therefore combined the two datasets into one.
We ran multiple exploratory analyses answering questions like:
-Which customers stay for long periods of time at the resorts?
-Customers from which countries generated the most revenue for the resorts?
-What were the meal preferences of the customers coming from within Europe?







After performing the exploratory analyses, we performed association rules mining and determined that previous guests who cancelled their booking were highly likely to cancel their bookings again for both locations. We determined rooms which should be a focus of the management for generating the highest revenue.

We fit the data into a linear model to find out the influence different variables had on the probability of cancelling a reservation.

Overall, this project had us dive into aggregating all the topics covered in the class to discover some very interesting behavioural patterns of guests based on the unstructured data provided.

Conclusion

I have been introduced to many new topics throughout my course in Applied Data Science, this portfolio helps me round off my learnings of two years at Syracuse University

I am learning more about web development and how to integrate machine learning models with website interfaces. I am excited to step into the market as a Data Scientist and putting my knowledge to use in the real world.