School of Information Studies
SYRACUSE UNIVERSITY

Final Milestone Portfolio

MSc Applied Data Science

Meichan Huang

mhuang01@syr.edu

Date due: March 20, 2024

Link to portfolio: https://github.com/mhgarrett/Meichan-Huang-SU-Applied-Data-Science-Portfolio-Project-Milestone-

Table of Contents

Final Portfolio Milestone Report for Msc Applied Data Science

Introduction

The Master's of Science of Applied Data Science at Syracuse University is designed to focus on the practical application of data science across various facets of enterprise operations, encompassing data acquisition and management, in-depth analysis, and the strategic communication necessary for informed decision-making.

My experience gained throughout this program equipped me with a solid foundation in both the theoretical underpinnings of data science and its real-world applications. I learnt to navigate the complexities of data mining, manage datasets, design and maintain databases, and data analysis to unveil valuable insights, and communicate these findings effectively to drive decision-making processes.

This portfolio showcases my ability to integrate the diverse knowledge and skills acquired, applying them to develop innovative solutions to operational challenges across different industries (airlines, medical, wine etc.). Specifically, proficiency in each fundamental aspect of data science showcased in this portfolio are as followed:

- Collect, store, and access data by identifying and leveraging applicable technologies
- Create actionable insight across a range of contexts (e.g. societal, business, political), using data and the full data science life cycle
- Apply visualization and predictive models to help generate actionable insight
- Use programming languages such as Rand Python to support the generation of actionable insight
- Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads)
- Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy)

Demonstrating mastery in data science requires more than understanding its principles; it demands the creative application of these principles to solve novel problems. In this portfolio, I demonstrate my ability to apply these principles in the following projects:

Term	Course Name	Project	Technical Skills	Tools
Fall 2022	IST 652: Scripting for Data Analysis & IST 644: Natural Language Processing	Wine review data analysis and text classification	Data visualization, Data analysis, Natural Language Processing, Word Clouds, Machine Learning	MangoDB, Pandas, NumPy, geopandas, sklearn, Seaborn, Matplotlib, NLTK, mapclassify
Spring 2023	IST 736: Text Mining	Airline Tweets Sentiment Analysis and Topic Modeling	Sentiment Classification, LDA Topic modeling, Natural Language Processing	NLTK, WordCloud, gensim, sklearn, spacy, and pyLDAvis
Spring 2023	IST 718: Big Data Analytics	Pneumonia X-Ray Image classification	Deep Learning, Neural Network, Image classification	Tensorflow, Kera, sklearn, Pandas, Numpy, and Matplotlib
Summer 2023	IST 722:Data Warehouse	FudgeWorld Data Warehouse Management	Database management, Data warehousing (ROLAP/MOLAP), ETL, Business Intelligence	SQL Server, SQL, SQL Server Integration Services (SISS), PowerBI
Fall 2023	IST 707: Applied Machine Learning	Credit Card Fraud Detection	Data preprocessing, SMOTE, Binary Classification	RStudio

Project 1: IST 652 & IST 644 - Wine Reviews Analysis and Classification

Project Description

The first project brought together two distinct sub-projects from IST 652: Scripting for Data Analysis and IST 644: Natural Language Processing courses I completed in the Fall of 2022. It was an exploration of the comprehensive data science lifecycle to foster actionable insights, from the wine consumers' perspectives. The aim was to predict wine scores by analyzing a blend of traditional statistical data and textual information, including origin, price, reviews from wine enthusiasts, and Twitter data. The project was structured in two segments, each with a different emphasis, showcasing the diverse data analytics skills I developed through IST 652 and IST 644.

The initial phase of the project was completed through IST652: Scripting for Data Analysis. The sub-project was marked by a strategic compilation of multiple external data sources to augment the primary dataset in Kaggle, which was scrapped from the Wine Enthusiast magazine during 2017-2020, containing 210,000 entries of wine reviews. Using Pandas package, the data was merged with the longitude and latitude dataset of the world geographic locations in Kaggle to visualize data distribution based on the locations.

This project enhanced my skills in thorough data exploration and analysis, incorporating traditional correlational analysis to examine the interplay between various attributes. The methodologies applied during this stage leveraged Python's robust libraries, specifically Pandas and NumPy, to facilitate data manipulation and analysis. The focus of the first part of the project was to uncover the relationships among data attributes and to understand the dataset's overall structure. Insights gleaned from this phase were aimed at data visualization and exploratory analysis, offering both graphical and statistical insights into the inherent patterns and trends within the wine review data. Key questions included identifying the top 10 most reviewed wine types and the most frequently reviewed wine regions. The following demonstrated one of such exploratory analysis done using the dataset to gain insights into the wine around the world, including the origins of the wine that were reviewed by the magazine (Figure 1).

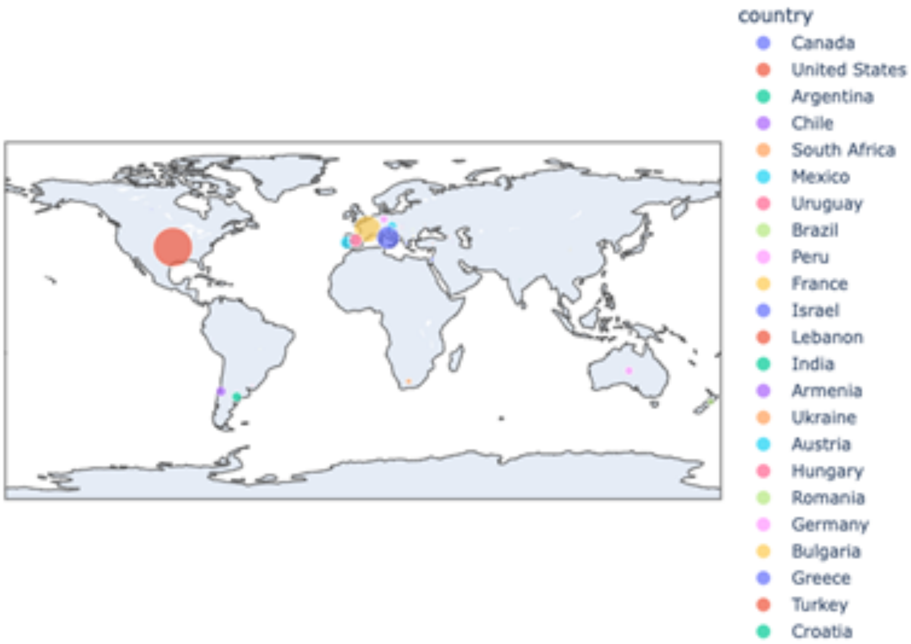


Figure 1. Top most reviewed wine origins by countries

Additionally, the project presented me with an opportunity to explore the relationship between wine prices and their ratings, yielding insights with significant business implications. Specifically, it revealed that wines with high ratings are not necessarily associated with higher prices. To uncover this, a regression analysis was conducted using the sklearn library, examining the correlation between wine prices and the points awarded. A noteworthy hurdle of this analysis was the creation of visualizations to convey meaningful insights effectively. This was particularly important given the data's characteristics: a narrow points range (80 - 100) contrasted with a vast, densely populated price range, mostly concentrated within the lower consumer price quartile. Consequently, the correlation model was refined to consider only wines priced at $100 or less, enabling a clearer comprehension of the relationship between price and quality.

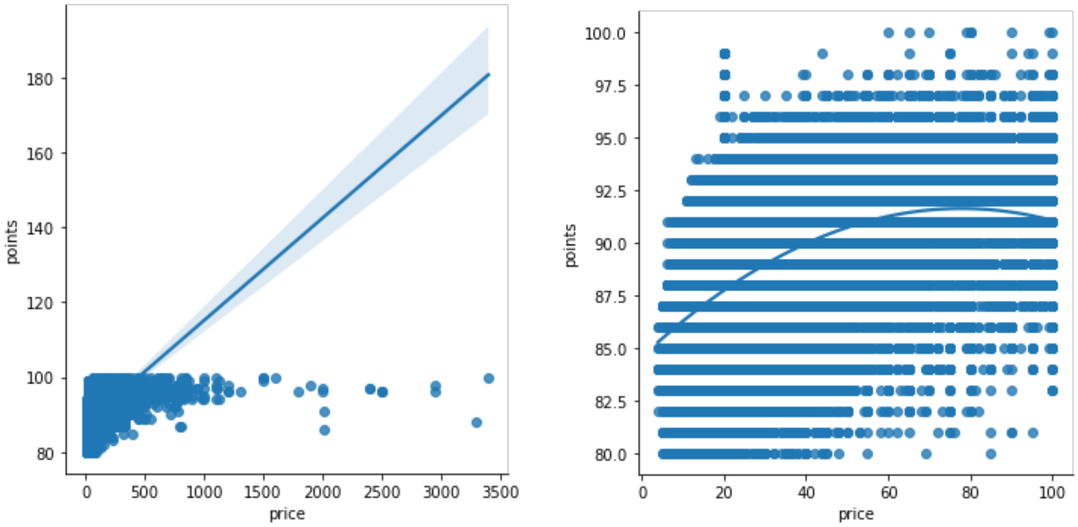


Figure 2: Scatter plots of price and points (raw and adjusted for price point)

Like any research projects I have conducted, this project had its constraints. The initial phase did not fully leverage the textual data present in the primary dataset. Acknowledging this limitation, I strategized to explore this rich textual content in the subsequent stage of my research, aiming to harness the full potential of the available data.

The second segment of the project, executed as part of IST 644: Natural Language Processing, built upon the foundational work of IST 652: Scripting for Data Analysis. In this phase, I capitalized on the textual data embedded within the wine reviews. The aim was to harness linguistic features from these reviews to predict outcomes such as high vs. low ratings, price brackets, and wine varietals. Collaborating with Nicholas Nguyen, we delved into natural language processing techniques. Our toolkit included tokenization and POS tagging via the NLTK library, extracting features using the bag-of-words model and bigrams, and implementing vectorization through TF-IDF scores. To bring the data to life, we created visual word clouds.

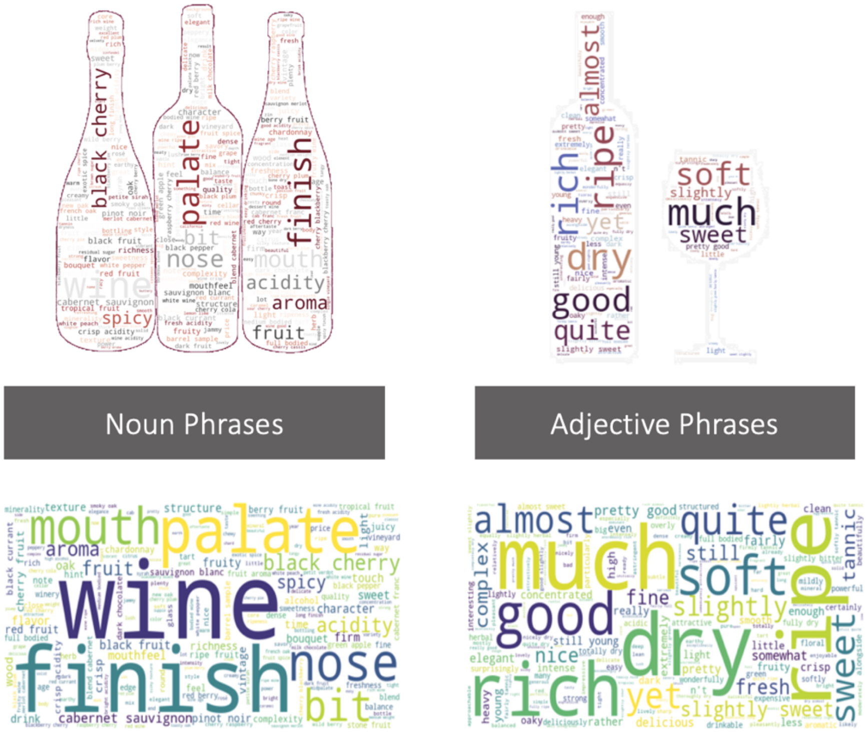
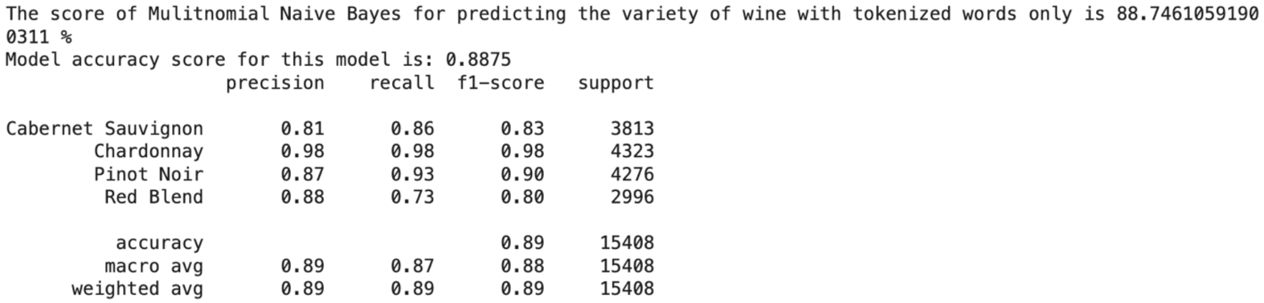


Figure 3. Word Clouds generated with different Part-of-Speech (POS) labels

In our final analysis phase, we engaged with text classification using the Multinomial Naïve Bayes algorithm, exploring the expanse of rich textual data to extract predictive insights. We experimented with various models to uncover patterns within the data, such as examining the predictive power of adjectives on wine varieties and analyzing the influence of tokenized words on wine classification. Among the models tested, it was the tokenized words from wine reviews that yielded the highest accuracy in predicting wine types (see Figure 4). This outcome indicates a tendency for reviewers to employ specific vocabulary when describing different wines, reflecting a distinctive linguistic pattern correlated with wine categories.



```
The score of Mulitnomial Naive Bayes for predicting the variety of wine with tokenized words only is 88.7461059190
0311 %
Model accuracy score for this model is: 0.8875
                    precision    recall  f1-score   support

Cabernet Sauvignon       0.81      0.86      0.83      3813
        Chardonnay       0.98      0.98      0.98      4323
        Pinot Noir       0.87      0.93      0.90      4276
         Red Blend       0.88      0.73      0.80      2996

          accuracy                           0.89     15408
         macro avg       0.89      0.87      0.88     15408
      weighted avg       0.89      0.89      0.89     15408
```

Figure 4. Multinomial Naïve Bayes model accuracy using tokenized words in wine reviews to predict wine varieties

Reflection and Learning Goals

The wine review project stands out as an important experience in my educational journey, marking my initial attempt into the data scientist path. This project was also crucial for my first programming experience, where I developed foundational data science skills from the ground up, with a particular focus on Natural Language Processing—a key aspect of my professional trajectory.

One of the most significant lessons from this project, which aligns with the learning goals of the program, was the realization that data analysis and data science projects are inherently cyclical, not linear. This iterative process often involves revisiting and refining each stage to achieve the most insightful results. Through this experience, I learned that each phase of analysis can shed new light and necessitate adjustments in strategy or technique. This project has not only equipped me with technical skills but has also instilled a mindset geared towards persistence and innovation—qualities that are indispensable for any aspiring data scientist.

Project 2: IST 736 - Airline Tweets Sentiment Analysis and Topic Modeling

Project Description

In the IST 736: Text Mining course instructed by Dr. Norma Grubb in Spring 2023, I enhanced my understanding of text mining fundamentals, including document representation, information extraction, text classification and clustering, and topic modeling. Teamed with Ryan Tervo, we analyzed a corpus of 14,640 tweets from Kaggle.com to gauge public sentiment toward six major airlines in the U.S. This project encompassed two types of advanced text-mining techniques that I have acquired through the course, namely sentiment classification and topic modeling.

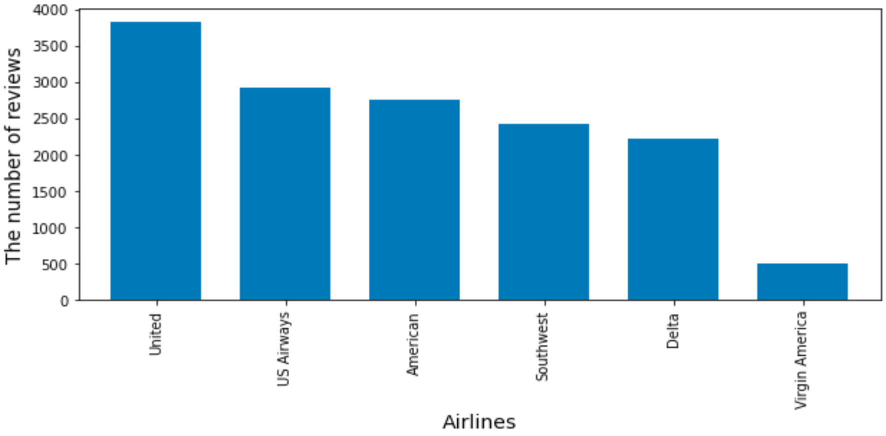


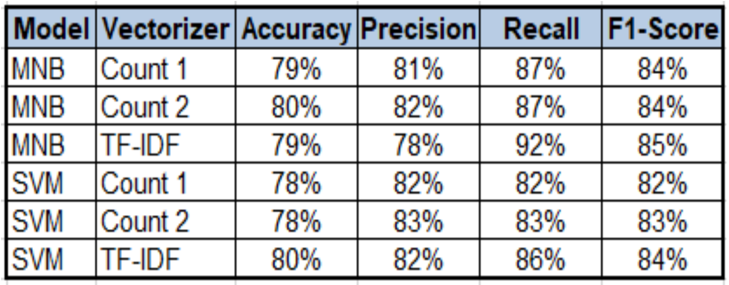
Figure 5. Tweet counts by airlines in the dataset

Through the use of different classification algorithms (naïve bayes and SVM), we tested different feature engineering and parameter tuning techniques to enhance the accuracy in classify sentiment in tweets automatically and practiced. In this analysis, different feature engineering techniques were deployed based on the structure of the tweet corpus, specifically, how to remove special characters, symbols, and stopwords, vectorize the tokenized data (i.e. binary vectorizer, count vectorizer, and TF-IDF vectorizer) and tune parameters to control for upper-case and number of n_grams, and maximum number of features to extract based on the TF-DIF scores etc, see examples of the parameters below:

Parameters	Setting
encoding	‘latin-1’
binary	False
min_df	5
smooth_idf	True
sublinear_tf	True
ngram_range	(1, 4)
max_features	2000

Table 1. Parameter setting example for sentiment classification

Then, six models were evaluated a second time in which the category labels were modified to be either ‘negative’ and ‘notNegative’. The ‘notNegative’ category was the combination of ‘positive’ and ‘neutral’ tweets. The following is a compilation of the results. Overall, the best accuracy was achieved using the SVM model with the TF-IDF vectorizer which improved the accuracy of the trivial model by 13.1%. The trivial (no information) model would predict every tweet ‘negative’ and achieve an accuracy of only 62.9%.



Model	Vectorizer	Accuracy	Precision	Recall	F1-Score
MNB	Count 1	79%	81%	87%	84%
MNB	Count 2	80%	82%	87%	84%
MNB	TF-IDF	79%	78%	92%	85%
SVM	Count 1	78%	82%	82%	82%
SVM	Count 2	78%	83%	83%	83%
SVM	TF-IDF	80%	82%	86%	84%

Table 2. Model results of sentiment classification

Additionally, we investigated the topic patterns of positive and negative tweets to gain insight into customers' perceptions of airlines. We wanted to gain a deeper understanding of the factors that drive sentiment in the airline industry for insights into customer preferences and opinions. Therefore, Latent Dirichlet Allocation (LDA) was employed as our topic modeling technique to decipher latent themes in the tweets, which is an approach that not only aligned with our project's goals but also enhanced our understanding of the data's underlying structure. To improve the clarity of our results, we integrated human interpretation with the algorithmic output, ensuring that the themes we presented were both accurate and resonant with human understanding. instead of relying on the matrices of log-likelihood and perplexity scores. For instance, the following is one of these selected negative tweets LDAVis topic modeling output based on the judgment of both modeling performance matrices and human judgment with three topic clusters.

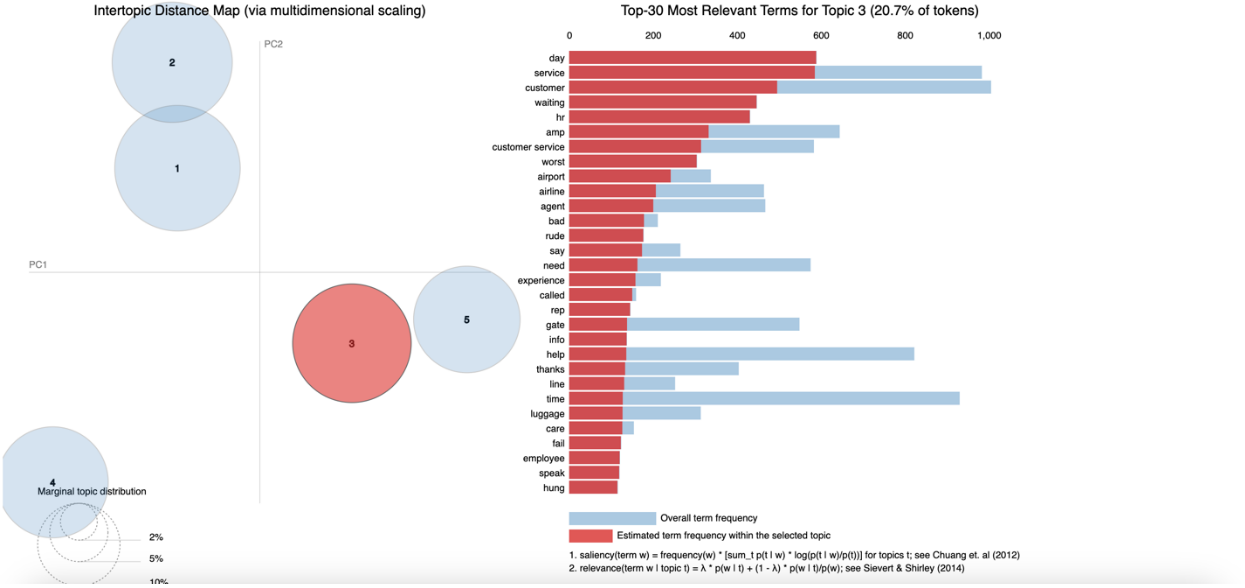


Figure 6. LDAVis output for 3 topic clusters in negative tweets and 30 most relevant terms

Advanced Python packages were employed to assist our analysis, for instance, one of the packages that genism library for text-preprocessing and topic-modeling, adding on to my skill set for text analysis using NLTK package for natural language processing.

Reflection and Learning Goals

I see this project as a demonstration of my deepen understanding of data scientist techniques as I progressed along in the program. A pivotal skill I acquired was the adept use of different feature engineering strategies tailored to the dataset's structure. This involved mastering the art of vectorizing tokenized data—choosing between binary, count, and TF-IDF vectorizers—and fine-tuning parameters to address nuances such as stopwords, case sensitivity, and minimum document frequency. This experience was instrumental in enhancing my practical knowledge of machine learning algorithms and their application in text analysis tasks.

This hands-on experience also increased my awareness of strategic decision-making in selecting the right tools for our analysis. This essential skill—evaluating the strengths and limitations of different technologies—was put to the test in our project. Through this project, what I learned was that data science, particularly text-mining for public sentiment and opinions, is not just about applying text mining algorithms; it was about weaving together machine efficiency with human intuition to create a comprehensive narrative from the data. This approach has equipped me with the necessary skills and mindsets to navigate text analysis tools to ensure their effective use in my future data science projects.

Project 3: IST 716 - Pneumonia X-Ray Image classification

Project Description

The final project of IST 716: Big Data Analysis was an attempt for me to embrace other areas of ML application – image classification with chest x-ray images from a Kaggle competition. Working with Matthew Pergolski and Shawn Anderson as a team, we explored the efficacy of deep Convolutional Neural Networks (CNNs) in distinguishing between normal and abnormal chest X-rays. As an exploration in the deep learning, we leveraged 5,863 chest x-rays from children aged one to five years old with each image categorized by its respective diagnosis, normal or pneumonia.

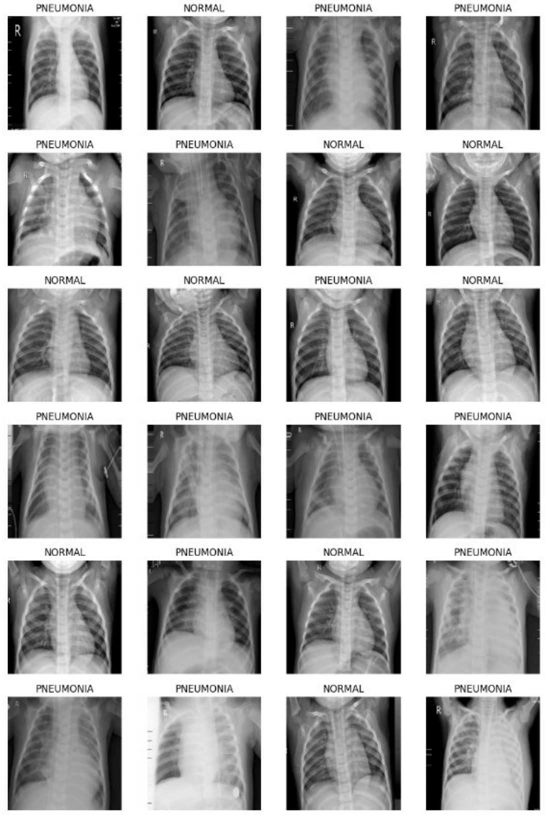


Figure 7. Chest X-ray data samples from the dataset

The highlights of this project was applying two distinct deep learning models, ResNet and VGG-16 for the image classification, each with various parameter tunings, which required using PyTorch to deploy these models. This was challenging as my first attempt for deep learning models. However, through research, we were able to complement PyTorch with Fastai, a high-level library designed, which simplified the building and training of machine learning models. To monitor our progress, we kept a tab on several metrics: the number of epochs (full cycles through the training data), training and validation losses (indicating how well the model was fitting the data), error rates (showing the frequency of incorrect predictions), and the time taken for each epoch. For instance, the following results demonstrated the matrices that we used to measure the performance of the models.

Model	**Accuracy**
RESNET18	84.78%
RESNET50	81.89%
RESNET152	85.58%

Table 3. Model Accuracy of RESNET models.

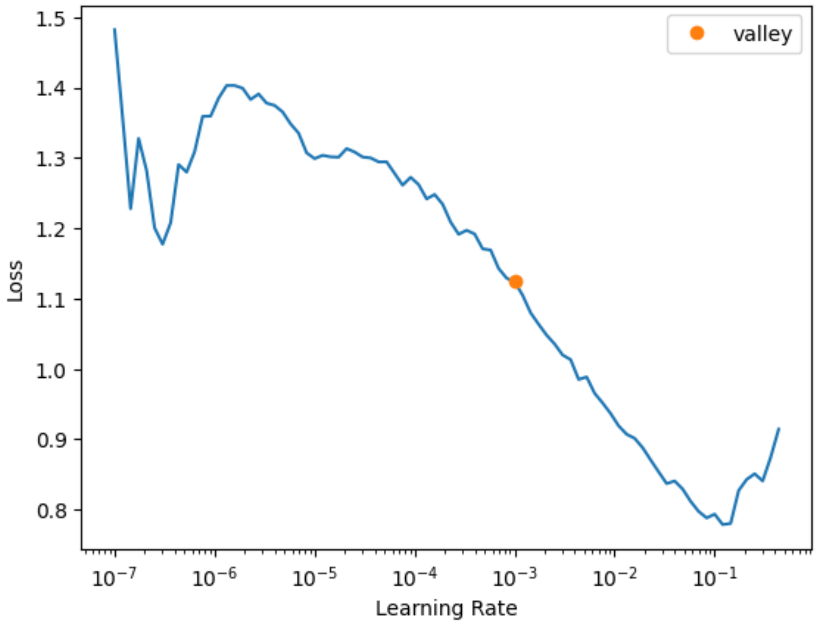


Figure 8. Learning Rate of RESNET model

Another example here is VGG 16 performance matrix output that showed the accuracy, loss and accuracy in one. It also showed the speed of learning in each epochs.

```
163/163 [==============================] - 163s 930ms/step - loss: 0.4182 - accuracy:
0.7947 - val_loss: 0.7801 - val_accuracy: 0.6250
Epoch 2/3
163/163 [==============================] - 159s 975ms/step - loss: 0.2603 - accuracy:
0.8928 - val_loss: 0.4858 - val_accuracy: 0.8125
Epoch 3/3
163/163 [==============================] - 160s 982ms/step - loss: 0.2140 - accuracy:
0.9160 - val_loss: 0.6476 - val_accuracy: 0.7500
```

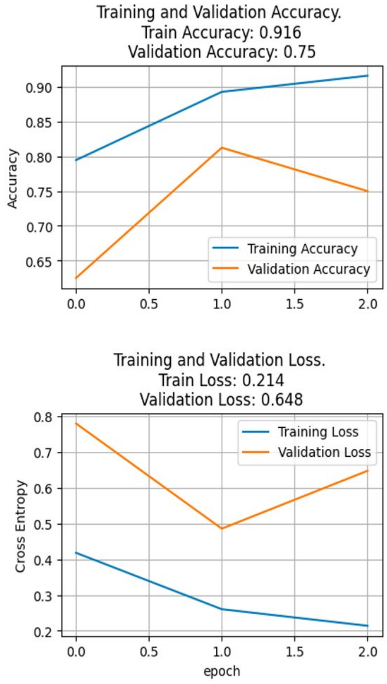


Figure 9. Training and Validation Accuracy and Loss in VGG16 models

Another highlight in the project was that, through researching in literature, I was able to utilize Grad-CAM (Gradient-weighted Class Activation Mapping) to overlay the chest X-rays when running VGG-16 models, which generates a heatmap underlining the significant regions or features in an image that the network primarily uses for prediction, allowing for easier visualization.

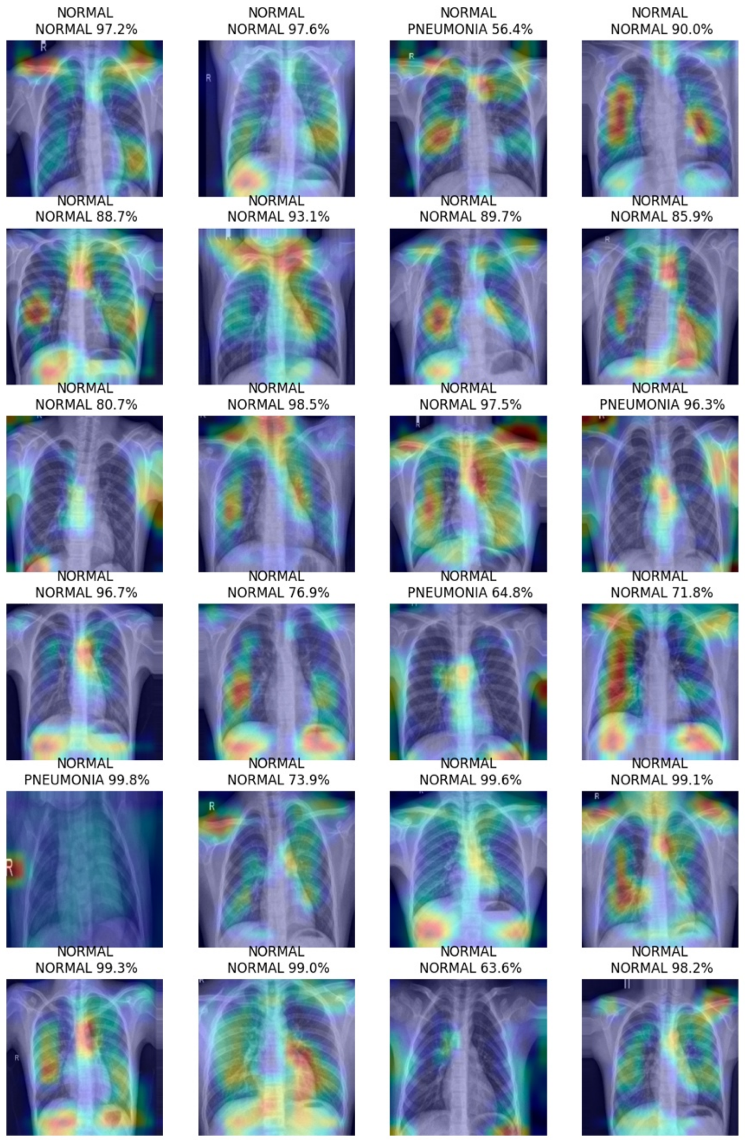


Figure 10. Grad-Cam overlay on the Chest X-ray on features used for prediction

Reflection and Learning Goals

This project was a break-through for me and also an eye-opener for me. It extended my data science toolkits in including practical machine learning applications, specifically in image classification. By applying advanced models like ResNet and VGG-16, each with unique parameter adjustments, I familiarized myself with the application of Convolutional Neural Network, utilizing python packages PyTorch and Fastai libraries, which embodied the program's objective of leveraging cutting-edge technologies for data analysis.

It is also a demonstration of my commitment to fostering advanced analytical skills and providing a foundation for evaluating and applying these skills to real-world challenges, as the program has taught me.

Project 4: IST 722 - FudgeWorld (Fudgemart and Fedgeflix) Data Warehouse Management

Project Description

In the IST722 Data Warehouse course instructed by Dr. Humayun Khan during Summer 2023, I acquired a holistic understanding of business intelligence (BI) and the pivotal role of data warehouse solutions in advancing organizational decision-making processes. This comprehensive course enabled me to acquire the technical aspects of data warehouses, offering me the knowledge to navigate through complex database constructs such as Operational Data Stores (ODS), Data Warehouses, and Data Marts. Furthermore, the course sharpened my skills in describing and applying various data integration approaches, including Extract, Transform, Load (ETL), Enterprise Information Integration (EII), and Enterprise Application Integration (EAI).

The final project for IST 722: Data Warehouse served as a comprehensive demonstration of such knowledge and skills acquired throughout the course. The knowledge and skills involved are essential for managing the seamless flow of data across various systems, ensuring both its quality and consistency. The project involved working with two simulated data warehouses, FudgeMart and FudgeFlex, which are analogous to Amazon and Netflix, respectively. These entities were presented as a merged company, setting the stage for our objective: to construct a BI solution from the ground up. Since both Fudgemart and Fudgeflix had relational databases before the merge and both companies made a large investment in data and data analysis. With the new merger management decided that a merged data warehouse was needed to allow data analysis from both companies as well as historical data from prior to the merger. Management wants the ability to do analysis of Fudgeworld sales while also being able to drill down into the specifics of which business line is contributing to the sales (i.e. fudgemart or fudgeflix).

Collaborating with Jake Conard, we constructed the FudgeWorld data warehouse using the ETL process, aiming to gaining insights into the sales of the merged company. This process required using different BI solutions, e.g. extracting the original data from SQL server, deploying SSIS to construct MOLAP cube and loading the data into a data warehouse to connect PowerBI to gain business insights of the company sales.

We started with the creation of a project charter to setting up our business objectives and shaping a comprehensive project schedule for our project. Then based on our business goals, we designed high-level dimensional modeling to outline the data structure of the data warehouse, i.e. what tables do we need to extract from the database for our analytic purposes.

To uphold data integrity and security, a crucial step involved extracting the data from its original database and creating a staged copy within the data warehouse. This step allows a controlled environment where data could be transformed and loaded without compromising its original state. To achieve this goal, we created the detailed-level dimensional modeling, which included the data's attributes and relationships, paving the way for the generation of an SQL script. This script was designed to facilitate the extraction of data into a data warehouse staged from the SQL server, where the original company data were stored.

SQL Server Integration Services (SSIS) was utilized to establish a connection with the SQL servers, enabling the extraction and loading of essential data tables into the Fudgeworld data warehouse, thereby facilitating comprehensive BI analysis aligned with our business objectives.

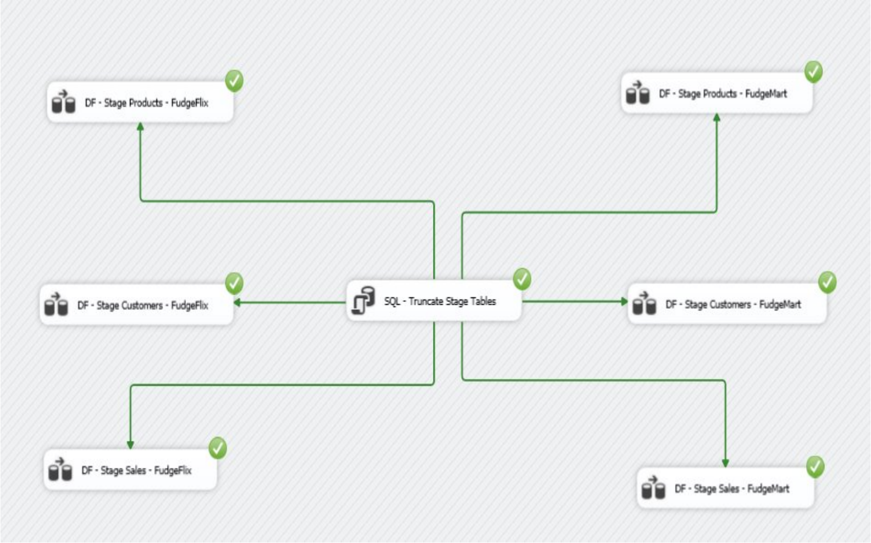

DF - Stage Products - FudgeFlix
DF - Stage Products - FudgeMart
DF - Stage Customers - FudgeFlix
SQL - Truncate Stage Tables
DF - Stage Customers - FudgeMart
DF - Stage Sales - FudgeFlix
DF - Stage Sales - FudgeMart

Figure 11: Extracting Data from SQL server

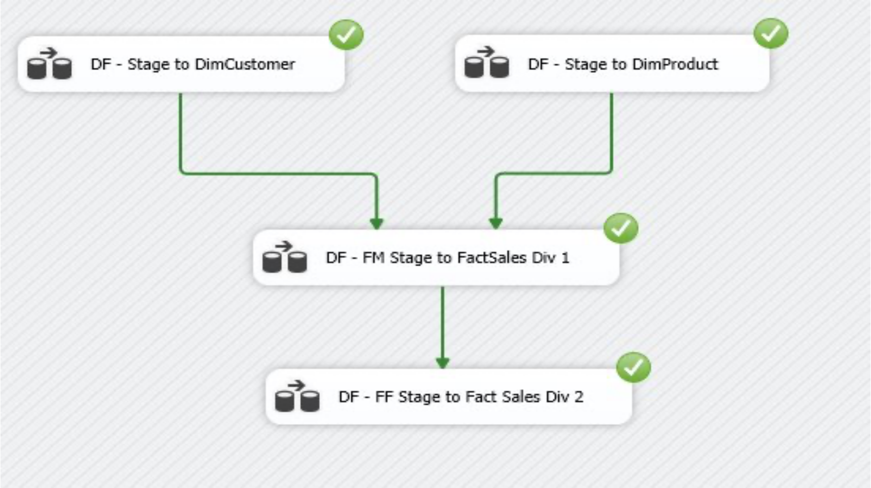
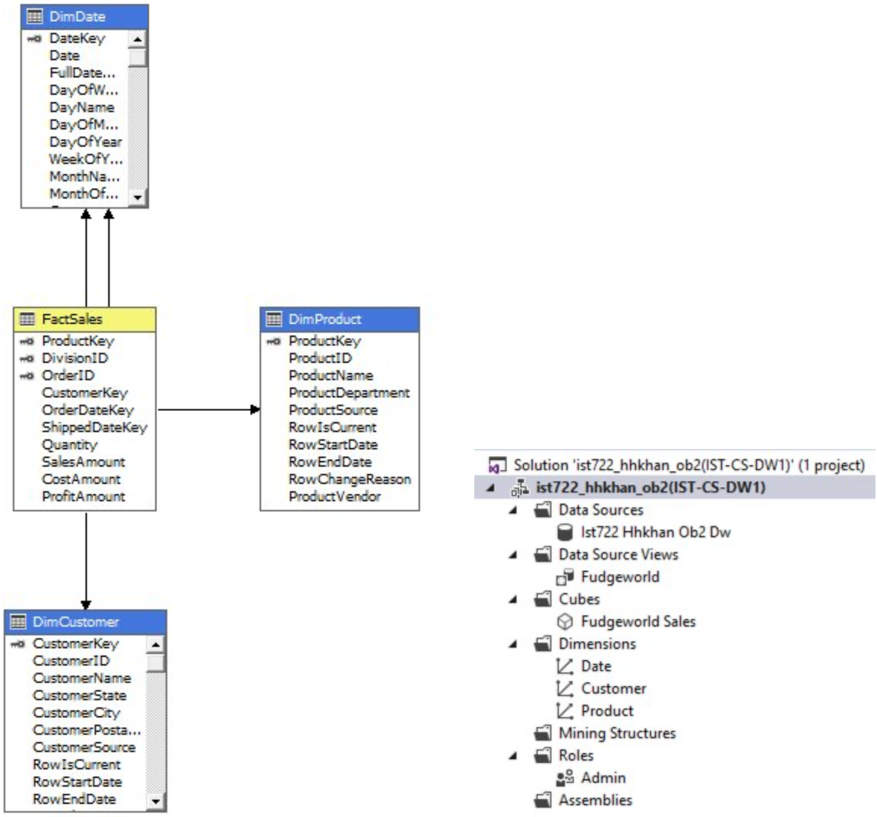

DF - Stage to DimCustomer
DF - Stage to DimProduct
DF - FM Stage to FactSales Div 1
DF - FF Stage to Fact Sales Div 2

Figure 12. Loading Data to Fact sheets in Data Warehouse

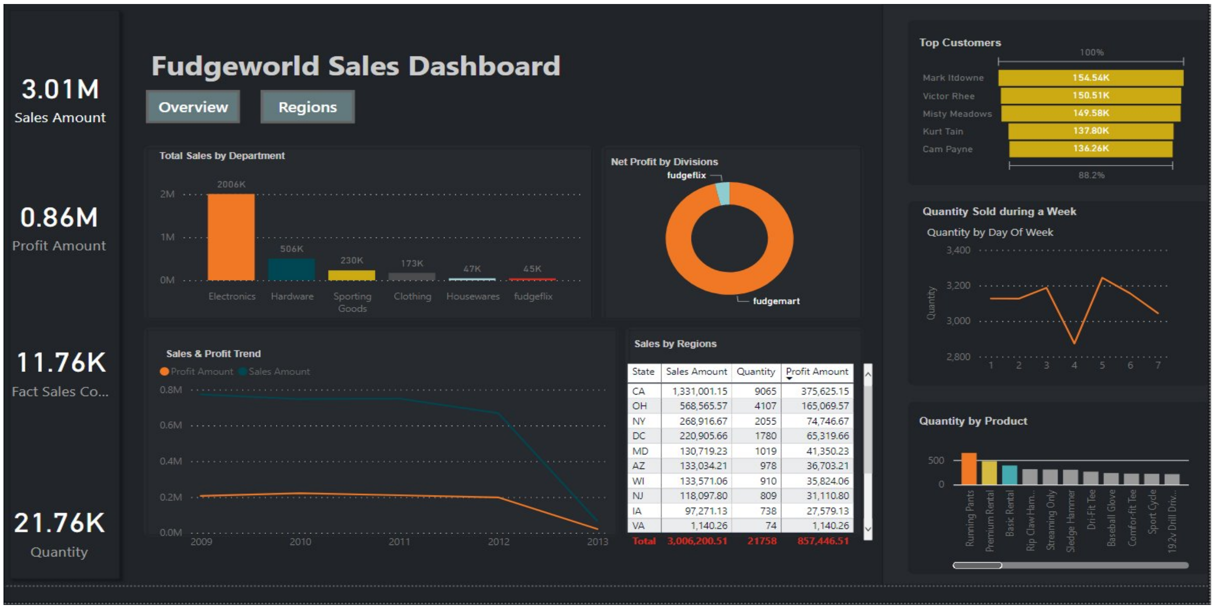
Then we incorporated the external sources (zipcodes and date variables) to create MOLAP cubes using SISS, enabling us to perform data analysis with enhanced efficiency and flexibility (Figure 1). The cubes facilitated the transformation processes, allowing for dynamic analysis and insights generation from the structured data.



```
Solution 'ist722_hhkhan_ob2(IST-CS-DW1)' (1 project)
  ist722_hhkhan_ob2(IST-CS-DW1)
    Data Sources
      Ist722 Hhkhan Ob2 Dw
    Data Source Views
      Fudgeworld
    Cubes
      Fudgeworld Sales
    Dimensions
      Date
      Customer
      Product
    Mining Structures
    Roles
      Admin
    Assemblies
```

Figure 13. MOLAP of FudgeWorld

Lastly, the data in the staged FudgeWorld data warehouse was loaded into PowerBI to create dashboards to gain insights into the business sales based on regions, months, and merchandise categories, as followed (Figure 14):



Sales by Regions

State	Sales Amount	Quantity	Profit Amount
CA	1,331,001.15	9065	375,625.15
OH	568,565.57	4107	165,069.57
NY	268,916.67	2055	74,746.67
DC	220,905.66	1780	65,319.66
MD	130,719.23	1019	41,350.23
AZ	133,034.21	978	36,703.21
WI	133,571.06	910	35,824.06
NJ	118,097.80	809	31,110.80
IA	97,271.13	738	27,579.13
VA	1,140.26	74	1,140.26
Total	3,006,200.51	21758	857,446.51

Figure 14. BI dashboard of FudgeWorld Sales

Reflection and Learning Goals

I believe that the final project for IST 722: Data Warehouse has encapsulated the essence of the program's goals: from data collection and management to insight generation and communication, with a strong emphasis on ethical considerations and practical applications. Reflecting on the process of completing the project, the technical application phase underscored the importance of programming proficiency and ethical data management practices. Specifically, utilizing SQL Server not only demonstrated my capability to structure and manage data with integrity but also instilled a deep appreciation for the ethical considerations that underpin data handling. The implementation of ETL processes using SSIS packages was a significant learning curve that honed my skills in data preparation and quality assurance. This experience illuminated the critical role of accuracy and usability in data science, reinforcing the necessity of meticulous data management. Furthermore, constructing an SSAS cube for analytics and developing BI dashboards and applications in Power BI was transformative to me. It not only solidified my technical skills but also enhanced my ability to communicate complex insights in an accessible manner to a diverse audience. This project through data warehouse construction, ETL processes, and BI development has been enlightening, challenging me to blend technical skills with ethical practices and effective communication to make data more meaningful and impactful.

Project 5: IST 707 - Credit Card Fraud Detection

Project Description

In IST707 Fall 2023 held by Dr. Jeremy Bolton, I expanded my understanding of how to extract meaningful knowledge from vast datasets. From the course work, I learned a spectrum of data mining methods, gained a solid theoretical foundation of machine learning algorithms, including data preparation, concept description, association rule mining, classification, clustering, evaluation, and analysis, were comprehensive. They offered me a holistic view of the data mining landscape, enabling me to solve real-world problems with practical applications.

For the final project, I chose to work independently with a large Kaggle dataset with 1,296,675 simulated credit card transactions, both legitimate and fraud, encompasses transactions from 1,000 customers across a network of 800 merchants and employed different classification algorithms to predict fraud cases. Since it is a binary classification problem, I have employed three popular algorithms in literature to solve this problem to identify the best performed algorithm. Specifically, traditional statistical method like logistic regression is known for their straightforward interpretability. Also, to tackle the increasing complexity of fraud patterns, I have also explored more advanced techniques. Specifically, I've utilized decision trees for their structured decision-making process and random forests to improve prediction accuracy. For this project, R studio was leveraged to conduct the data preprocessing and analysis and deploy subsequent machine learning algorithms.

One of the paramount challenges encountered in this project was data preparation, a task that provided significant learning opportunities. The dataset mirrored the real-world imbalance between fraud and non-fraud classes. To address this, two data conditions were scrutinized: the original dataset with its inherent class imbalances and a version processed using the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE, an approach that generates synthetic instances instead of merely duplicating the minority class, was instrumental in forging a more balanced dataset. This technique functions by creating synthetic examples through interpolation between a minority class observation and its nearest neighbors, thereby achieving a more equitable distribution of non-fraudulent and fraudulent cases for training purposes. Utilizing the SMOTE function led to a recalibrated dataset comprising 30,024 non-fraud and 22,518 fraud cases, effectively poised for algorithmic analysis. These two conditions were tested subsequently with all three algorisms.

To further optimize the performance of machine learning models in the face of numerous categorical variables and substantial variance across numerical data scales, a rigorous standardization process was undertaken. This involved normalizing the numeric data ranges and converting categorical data into a uniform format. A pivotal data transformation strategy I utilized was target encoding. This method proved invaluable for managing the high dimensionality of certain categorical variables by assigning new values to each category. These values were derived based on the category's correlation with the target variable, quantified as the proportion of the target outcome's occurrences within the category's total count. Target encoding is especially beneficial for tree-based algorithms such as decision trees and random forests, as it effectively addresses the "curse of dimensionality" that plagues one-hot encoding techniques.

Furthermore, to normalize numerical data with wide-ranging values, logarithmic transformation was applied. For example, the 'amount' attribute exhibited an extensive range, prompting the use of logarithmic scaling. This transformation is adept at compressing the spread of larger values while proportionately expanding the gaps between smaller figures, thus ensuring a more uniform data distribution conducive to model training and analysis.

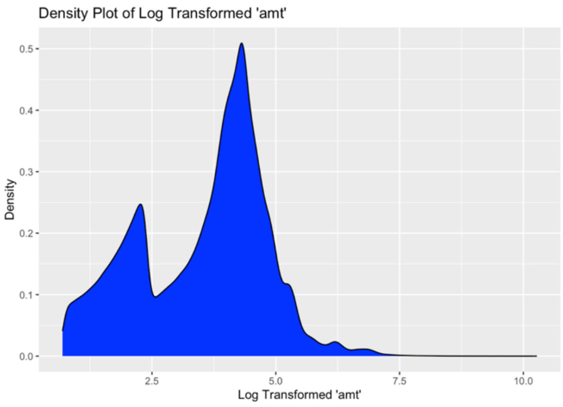


Figure 15. Log Transformation of Amount Attribute

Additionally, due to the high dimensionality of the dataset amongst the attributes, Principal Component Analysis (PCA) were carried out, however, this step did not successfully reduce the dimensionality, as the formed clusters did not significantly reduce the number of attributes within the dataset.

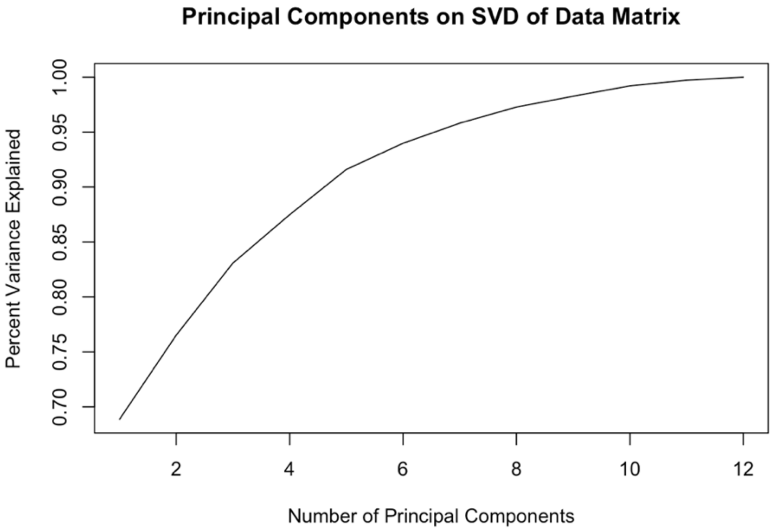
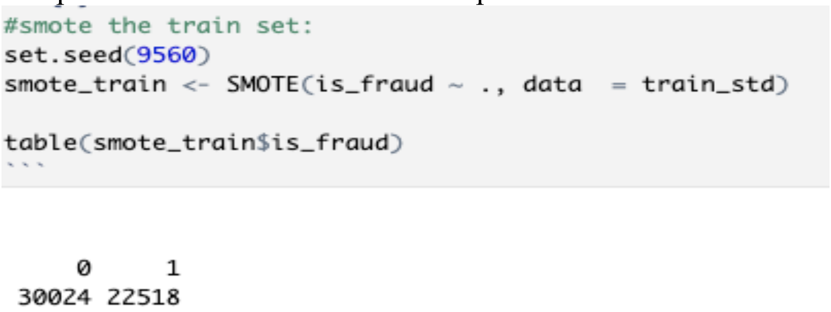


Figure 16. PCA results from the dataset

To effectively test the different data manipulation and tuning of the model, six conditions were set up and tested, Imbalanced v.s. SMOTE data using logistic regression, imbalanced and SMOTE data using Decision Tree, and imbalanced and SMOTE data using Random Forest. Another challenge posed in this project was, considering the dataset's characteristics, standard metrics such as accuracy may not be sufficient for model evaluation. Therefore, alternative metrics, including precision (also known as specificity), recall (also known as sensitivity), F1 scores, and the Area Under the Receiver Operating Characteristic Curve (AUC), were employed to provide a more comprehensive assessment of model performance.



~~~
#smote the train set:
set.seed(9560)
smote_train <- SMOTE(is_fraud ~ ., data  = train_std)

table(smote_train$is_fraud)
```
~~~

```
    0     1 
30024 22518 
```

Figure 17. R code for SMOTE sampling

The results showed that, comparing all six models, the Random Forest model using SMOTE data stands out as the most effective in fraud detection among the three tested algorithms, with a good balance of accuracy, recall, and a reasonable rate of false positives. Also, several attributes have shown to have strong indication to fraudulent activities. a strong indication to fraud, credit card fraud tends to concentrate in certain transaction categories such as shopping (both at physical locations and online) and grocery purchases. Patterns in fraudulent activities also vary by age, with spikes among those aged 35 and 50-55, and timing, with a higher incidence in early months of the year, particularly in May, and during weekends and late nights. Moreover, states with large populations and economic significance, like New York and California, report higher levels of fraud, possibly reflecting their dense and affluent environments. However, gender is not a strong indicator of fraud in this dataset.

Reflection and Learning Goals

Reflecting on the completion of the IST707 Fall 2023 project, it is clear how this project has solidified my understanding of data science principles and their application to complex real-world problems. The process of understanding, visualizing, and preprocessing a large dataset to increase the efficiency and accuracy in various classification algorithms was a huge endeavor to enhance my skill sets as a data scientist. It not only tested my technical skills but also pushed me to think critically about what entails being a data scientist, that is, 99% of the work involves data cleaning and preparation.

The challenge of addressing data imbalance through techniques like SMOTE was particularly illuminating. This part of the project was a stark reminder of the crucial role that data preparation plays in the broader context of data science, reinforcing that accurate analysis stems from diligently curated data.

Conclusion

To conclude, reflecting on my personal journey through the MSc Applied Data Science Program, I have witnessed my transformation from a novice in programming and technical knowledge to a future data scientist capable of informed decision-making based on the business objectives, the data provided, and the available technical tools at hand. I have not only gained extensive hands-on experience in collecting, storing, and accessing data, but also the ability to create actionable insights that are applicable across diverse contexts including business and societal.

These projects included in this milestone project have allowed me to practice in engaging with the full data science lifecycle, applying both visualization techniques and predictive models to cultivate insights that inform and inspire decision-making. My growing proficiency in programming languages such as R and Python has enabled me to fulfil this task, enabling me to not just understand but also to manipulate data in ways that reveal its underlying stories.