Syracuse University
School of Information Studies

Master of Science
Applied Data Science

Portfolio Milestone

Meichan Huang

mhuang01@syr.edu

March 20, 2024

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

Introduction: Project Milestones

The Master's of Science of Applied Data Science at Syracuse University is designed:

- Focus on practical application of data science across various facets of enterprise operations
- Gained a solid foundation in both theoretical underpinnings and real-world applications
- Showcases ability to integrate diverse knowledge and skills to develop innovative solutions

The following five projects demonstrated the skills developed in the program:

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	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

Learning Objectives within the Program

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)

Project 1: Wine Reviews Data Analysis and Text Classification

IST 652: Scripting for Data Analysis

IST 664: Natural Language Processing

Link to the project: https://github.com/mhgarrett/Meichan-Huang-SU-Applied-Data-Science-Portfolio-Project-Milestone-/tree/5bc88959660d050866e4397fe289a019c9df0118/Project%201%3A%20Wine%20review%20data%20analysis%20and%20text%20classification

	country	description	points	price	province	variety	tokenized_words
0	Italy	Aromas include tropical fruit, broom, brimston...	87	20.0	Sicily & Sardinia	White Blend	[aromas, include, tropical, fruit, broom, brim...
1	Portugal	This is ripe and fruity, a wine that is smooth...	87	15.0	Douro	Portuguese Red	[ripe, fruity, smooth, structured, firm, tanni...
2	US	Tart and snappy, the flavors of lime flesh and...	87	14.0	Oregon	Pinot Gris	[tart, snappy, lime, flesh, rind, dominate, gr...
3	US	Pineapple rind, lemon pith and orange blossom ...	87	13.0	Michigan	Riesling	[pineapple, rind, lemon, pith, orange, blossom...
4	US	Much like the regular bottling from 2012, this...	87	65.0	Oregon	Pinot Noir	[much, like, regular, bottling, comes, across,...

	geometry	Unnamed: 0	country	description	country_code	latitude	longitude
MULTIPOLYGON (((-122.84000 49.00000, -122.9742...		14	Canada	399	CA	56.130366	-106.346771
MULTIPOLYGON (((-122.84000 49.00000, -120.0000...		0	United States	86678	US	37.090240	-95.712891
MULTIPOLYGON (((-68.63401 -52.63637, -68.25000...		7	Argentina	5587	AR	-38.416097	-63.616672
MULTIPOLYGON (((-68.63401 -52.63637,		5	Chile	6068	CL	-35.675147	-71.542969

	Datetime	User	Location	Followers	TotalTweets	Likes
0	1669334354000	markreinoso	Senoia, GA	101	2595	3
1	1669334280000	jacy1273		65	317	0
2	1669334273000	Dedric84P		12	4553	0
3	1669333828000	InfoAuMax1		80	10736	0
4	1669333657000	bonesondisplay	Pennsylvania, USA	1830	17118	0

Overview

Data:

- 130k wine reviews from 2017 from Kaggle
- 80K wine reviews from 2017 – 2020 from Kaggle
- World Country Longitude and Latitude data from Kaggle
- 10k wine twitters with #wine, #winereview from 11/14–11/24/2022 scrapped using snscrape

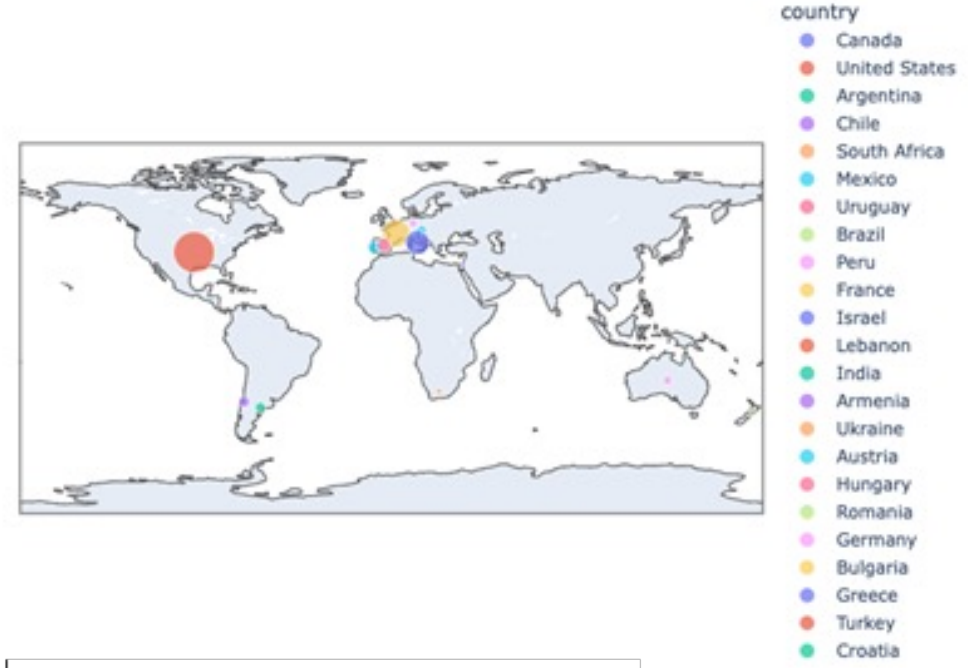
Techniques:

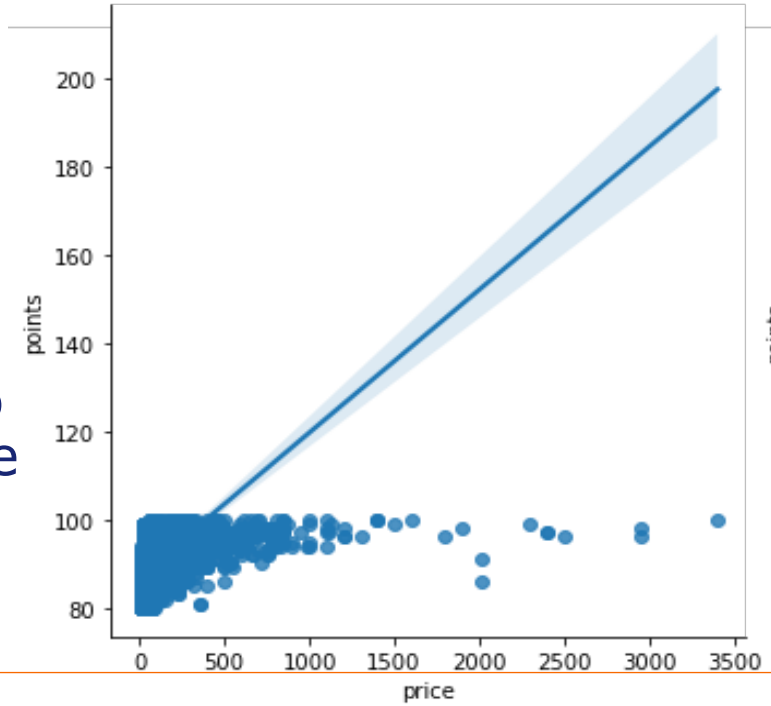
- Python packages
- Data visualization
- Data analysis
- Text classification

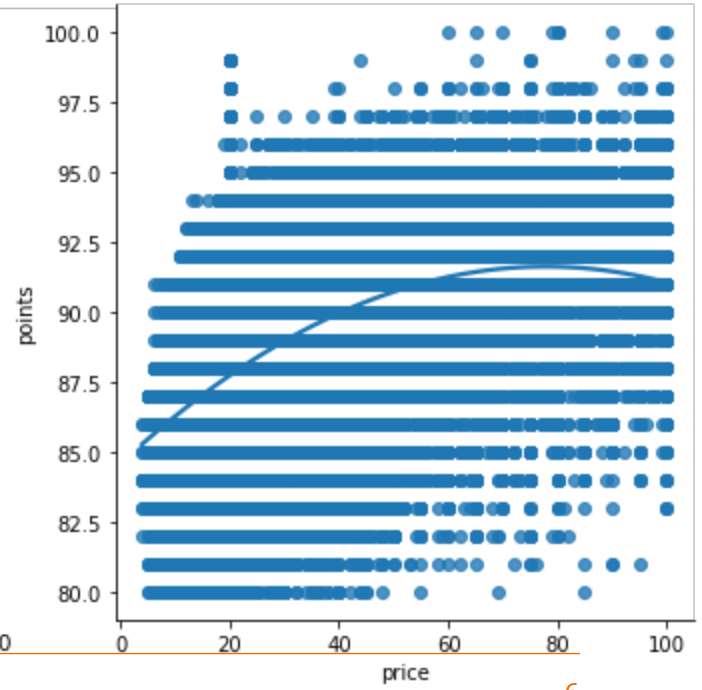
•Exploration of the comprehensive data science lifecycle
•Predict wine scores by analyzing traditional statistical data and textual information
•Structured in two segments, showcasing diverse data analytics skills

Project 1: Phase I (IST 652)

- Compiled multiple external data sources to augment primary datasets
- Utilized Python's libraries (Pandas, NumPy) for data manipulation and analysis
- Uncovered relationships among data attributes and overall structure
- Conducted regression analysis to examine correlation between wine prices and ratings







Project 1: Phase 2 (IST644)

- Capitalized on textual data embedded within wine reviews
- Employed natural language processing techniques (tokenization, POS tagging, vectorization)
- Created visual word clouds and implemented Multinomial Naïve Bayes algorithm for text classifications to understand specific vocabulary wine reviewers employ for varieties of wine

Noun Phrases

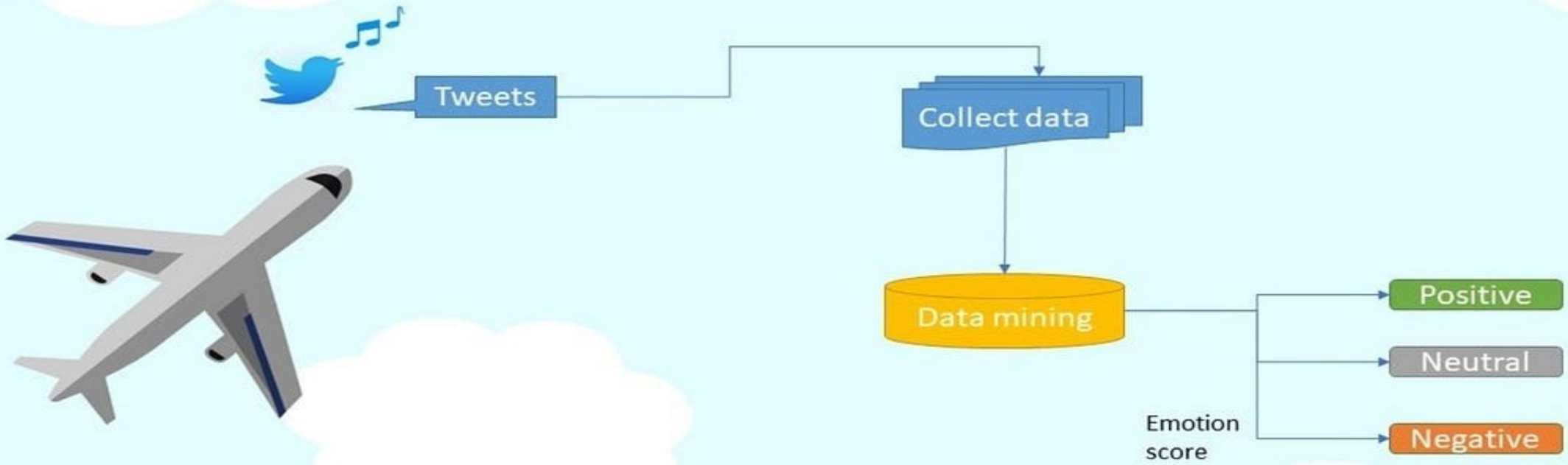
Adjective Phrases

The score of Mulitnomial Naive Bayes for predicting the variety of wine with tokenized words only is 88.74610591900311 %
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

Reflection of Learning Goals

- Developed foundational data science skills with a focus on Data Visualization, Data Analysis, and Natural Language Processing
- Learned that data analysis and data science projects are cyclical, not linear
- Equipped with technical skills and a mindset geared towards persistence and innovation



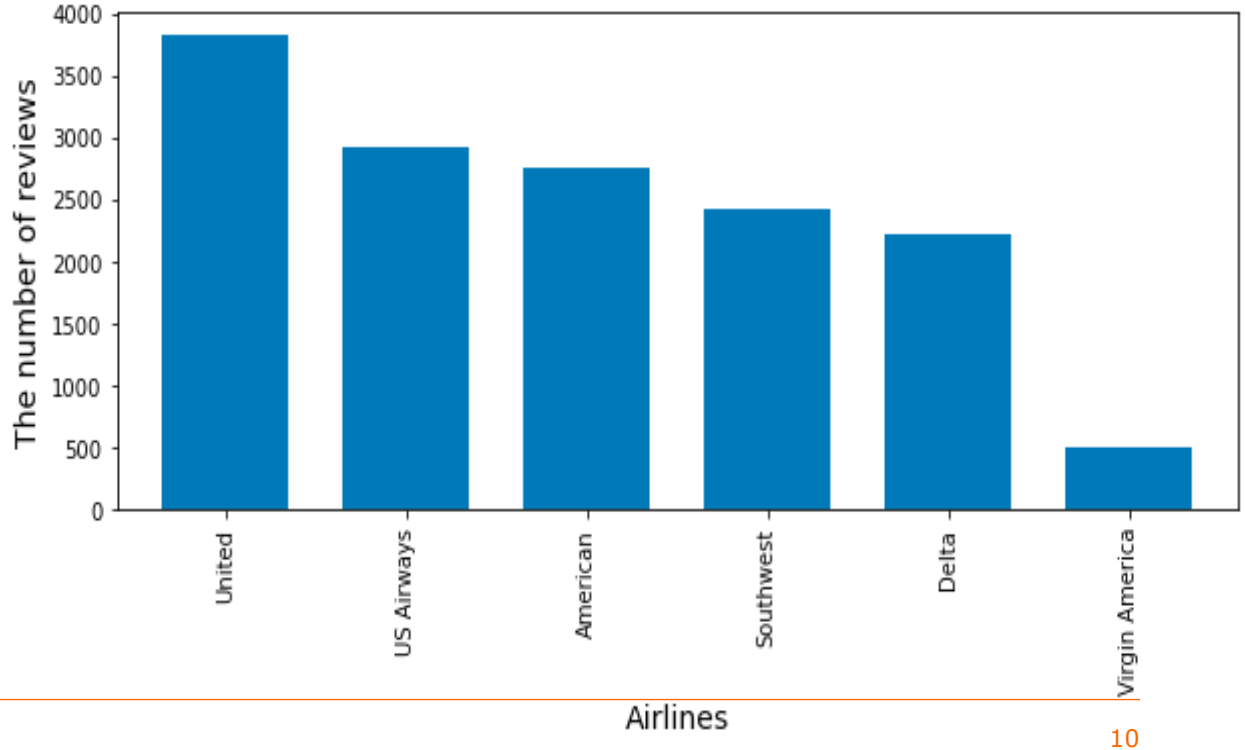
Project 2: Airline Tweets Sentiment Analysis and Topic Modeling

IST 722: Text Mining

Link: https://github.com/mhgarrett/Meichan-Huang-SU-Applied-Data-Science-Portfolio-Project-Milestone-/tree/8ada042d7762410e43fed4e0dd73be5558c75ad7/Project%202%3A%20Airline%20Tweets%20Sentiment%20Analysis

Overview

- Analyzed a corpus of 14,640 tweets to gauge public sentiment toward six major U.S. airlines
- Employed two types of advanced text-mining techniques:
  - Sentiment classification (negative, neutral, positive)
  - Latent Dirichlet Allocation (LDA) topic modeling



Sentiment Classification

- Deployed different vectorization methods (binary, count, and TF-IDF)
- Utilized various feature engineering and parameter tuning techniques (# cross-validation folds)
- Tested different classification algorithms (naïve bayes and SVM)

Parameters	Setting
encoding	'latin-1'
binary	False
min_df	2
max_df	1500
ngram_range	(1, 3)

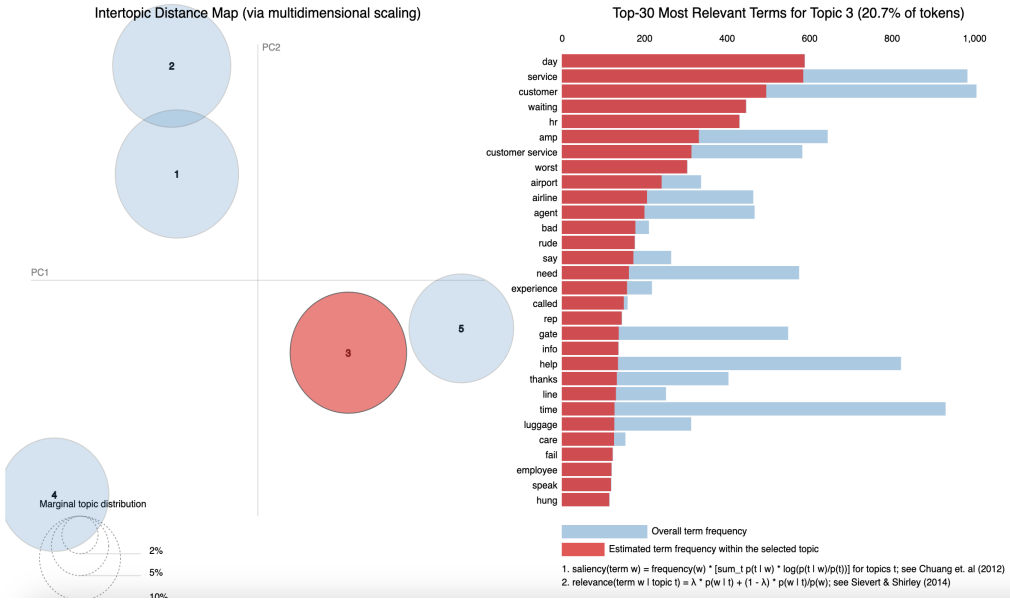
Parameters	Setting
encoding	'latin-1'
binary	False
min_df	5
smooth_idf	True
sublinear_tf	True
ngram_range	(1, 4)
max_features	2000

Model	Vectorizer	Accuracy	Precision	Recall	F1-Score
MNB	Count 1	75%	80%	88%	84%
MNB	Count 2	75%	79%	88%	84%
MNB	TF-IDF	74%	74%	96%	83%
SVM	Count 1	73%	82%	83%	83%
SVM	Count 2	74%	82%	84%	83%
SVM	TF-IDF	76%	81%	89%	85%

Topic Modeling

- Investigated topic patterns of positive and negative tweets
- Employed LDA as the topic modeling technique to decipher latent themes
- Integrated human interpretation with algorithmic output for accurate and resonant results: How many topic clusters are more reasonable based on human judgement vs. what modeling performance matrices are better?

Topic	Top 10 words in negative comments
1	time, hold, hour, agent, airline, aa, minute, worst, phone, wait
2	hour, plane, delayed, bag, late, delay, waiting, gate, hr, time
3	cancelled, customer, service, flightled, help, cancelled flightled, customer service, day, need, amp



Reflection of Learning Goals

Acquired adept use of feature engineering strategies tailored to dataset's structure

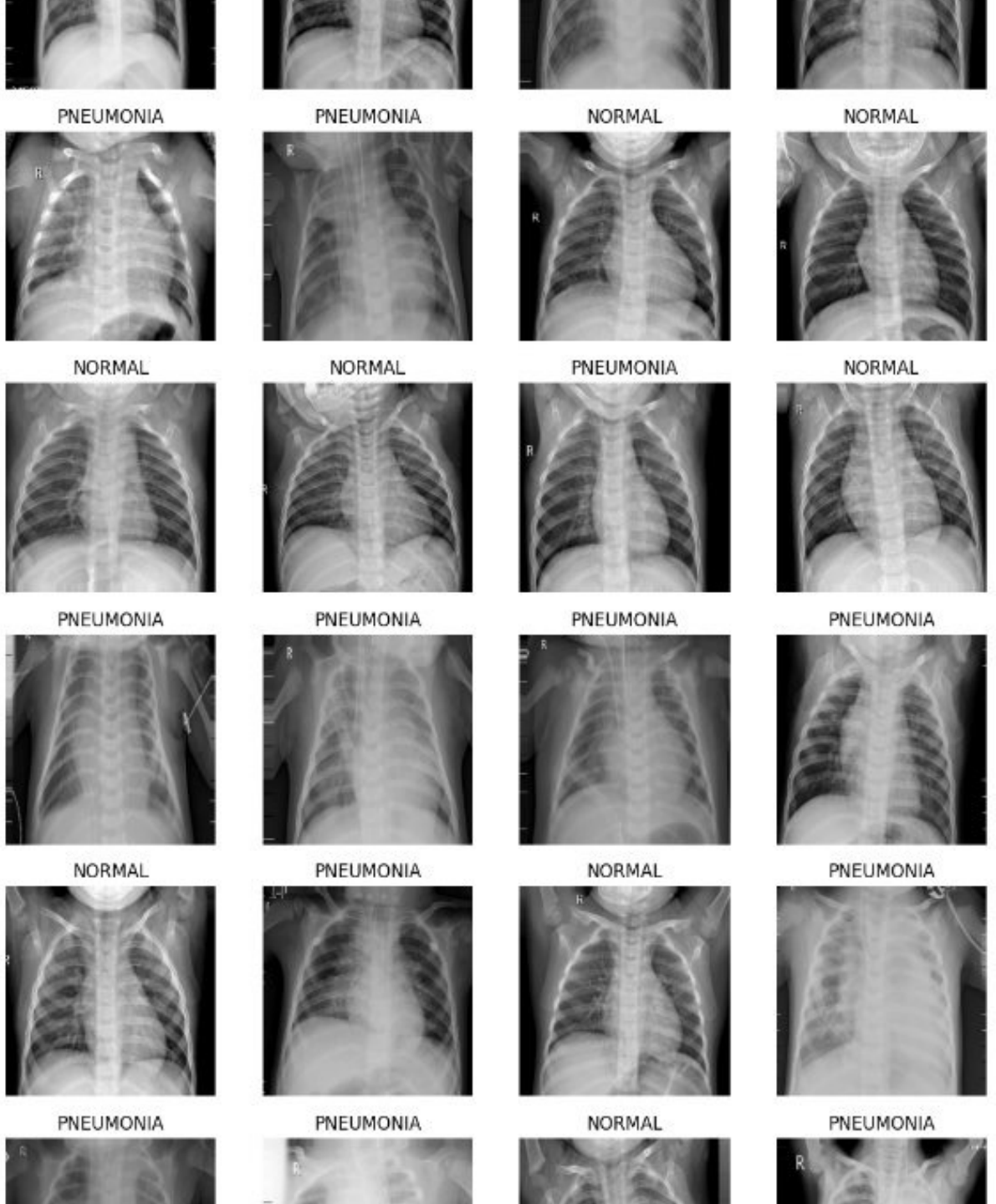
Enhanced practical knowledge of machine learning algorithms and their application in text analysis

Learned the importance of weaving together machine efficiency with human intuition

Project 3: Pneumonia X-Ray Image Classification

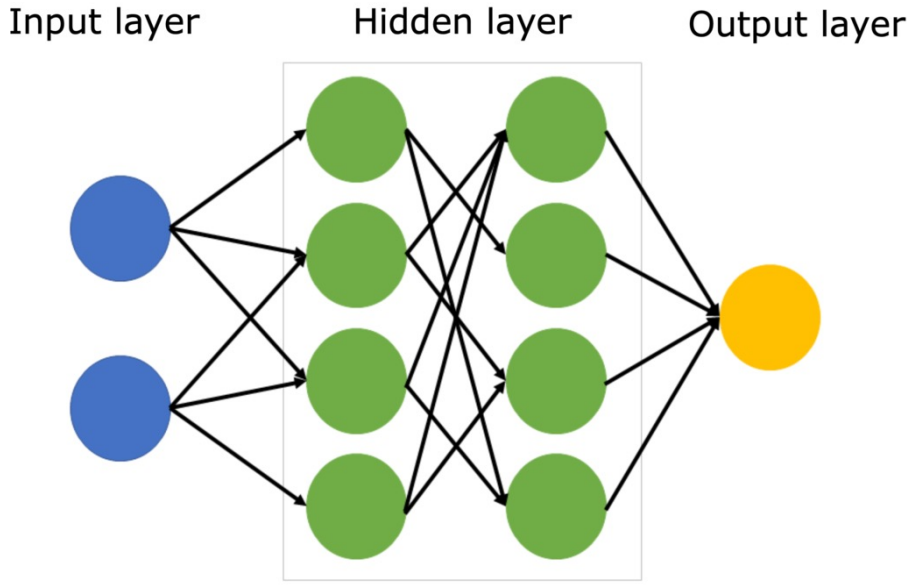
IST 718: Big Data Analytics

Link: https://github.com/mhgarrett/Meichan-Huang-SU-Applied-Data-Science-Portfolio-Project-Milestone-/tree/f5d9fd68d070a46ff94c3353477e0c297a0a5d41/Project%203%3A%20Pneumonia%20Chest%20X-ray%20Image%20Classification

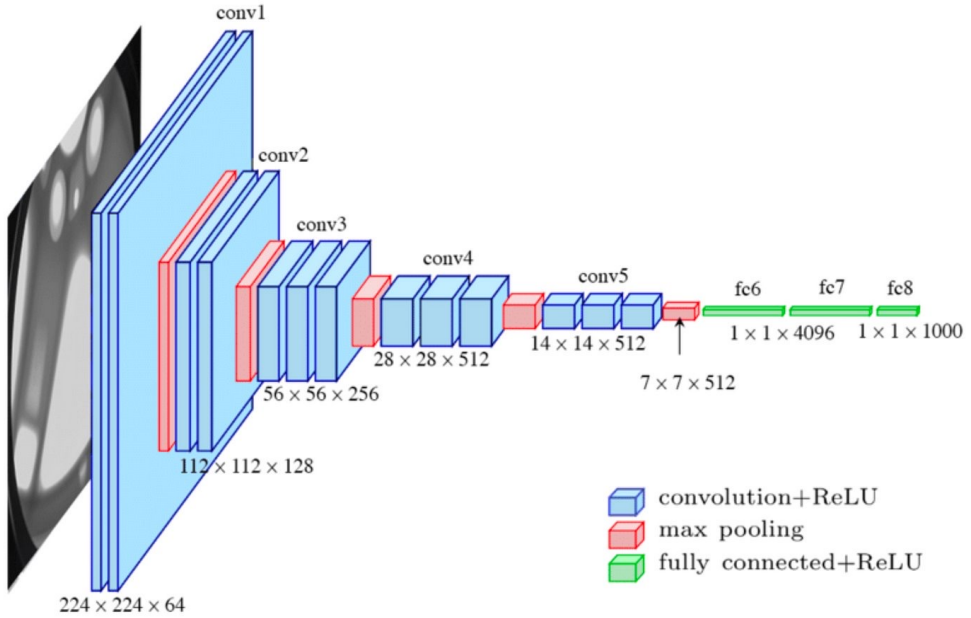


Overview

- Explored the efficacy of deep Convolutional Neural Networks (CNNs) in distinguishing between normal and abnormal chest X-rays
- Utilized 5,863 chest X-rays from children aged one to five years old



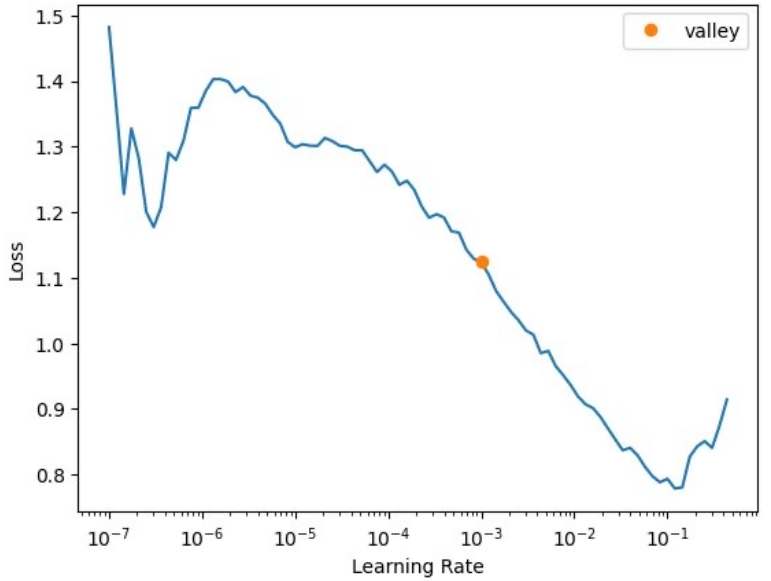
Artificial neural networks

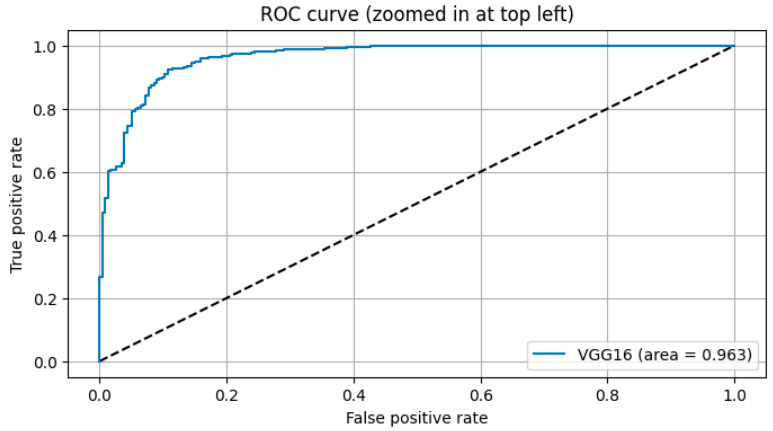


Models for Images Classification & Model Tuning

- Applied two distinct deep learning models: ResNet and VGG-16
- Employed PyTorch and Fastai libraries for model building and training
- Monitored metrics such as epochs, training and validation losses, error rates, and time taken
- Utilized Grad-CAM (Gradient-weighted Class Activation Mapping) to overlay chest X-rays
- Generated heatmaps underlining significant regions or features for prediction
- Allowed for easier visualization of the model's decision-making process

Model	Accuracy
RESNET18	84.78%
RESNET50	81.89%
RESNET152	85.58%
RESNET18 (Learning Rate Optimized)	85.26%





Reflection of Learning Goals

- Extended data science toolkit to include practical machine learning applications in image classification
- Familiarized with the application of Convolutional Neural Networks using PyTorch and Fastai libraries
- Demonstrated commitment to fostering advanced analytical skills and applying them to real-world challenges

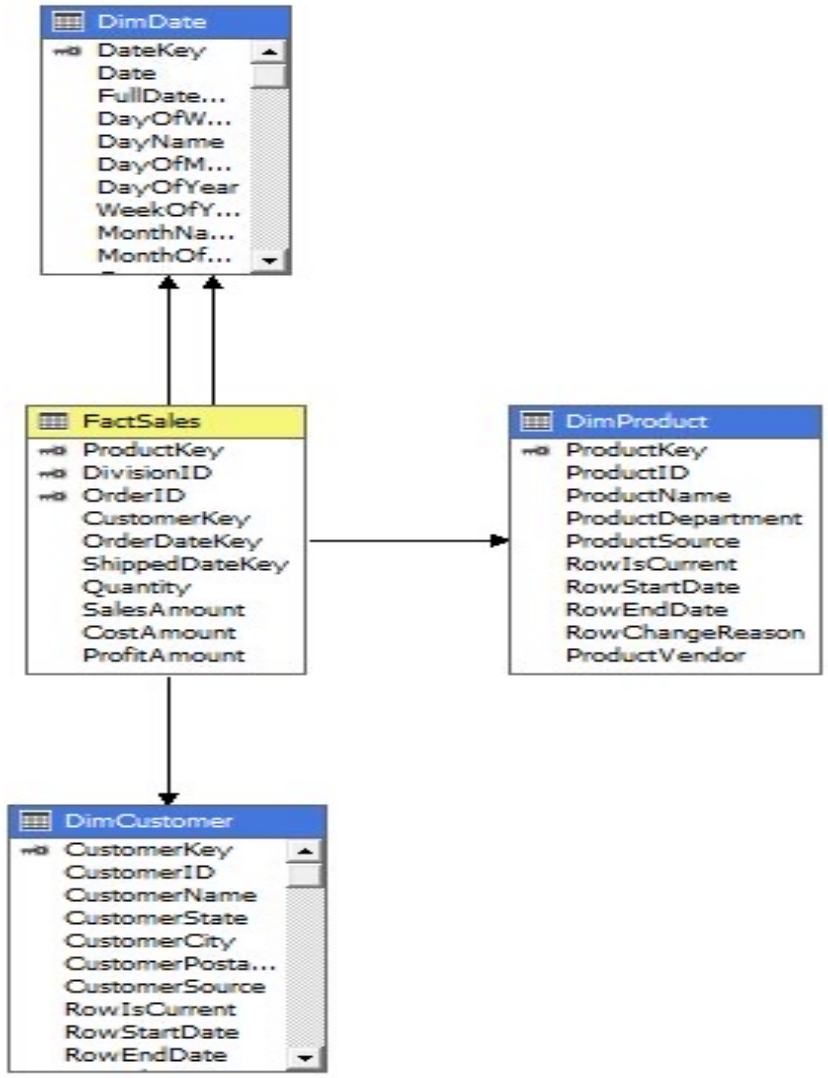
FudgeWorld (Fudgemart and Fedgeflix) Data Warehouse Management

IST 722: Data Warehouse

Link: https://github.com/mhgarrett/Meichan-Huang-SU-Applied-Data-Science-Portfolio-Project-Milestone-/tree/689f434cc268a996ca10f81ea5022f6f3b89d7ba/Project%204%3A%20Fudgeworld%20Data%20Warehousing

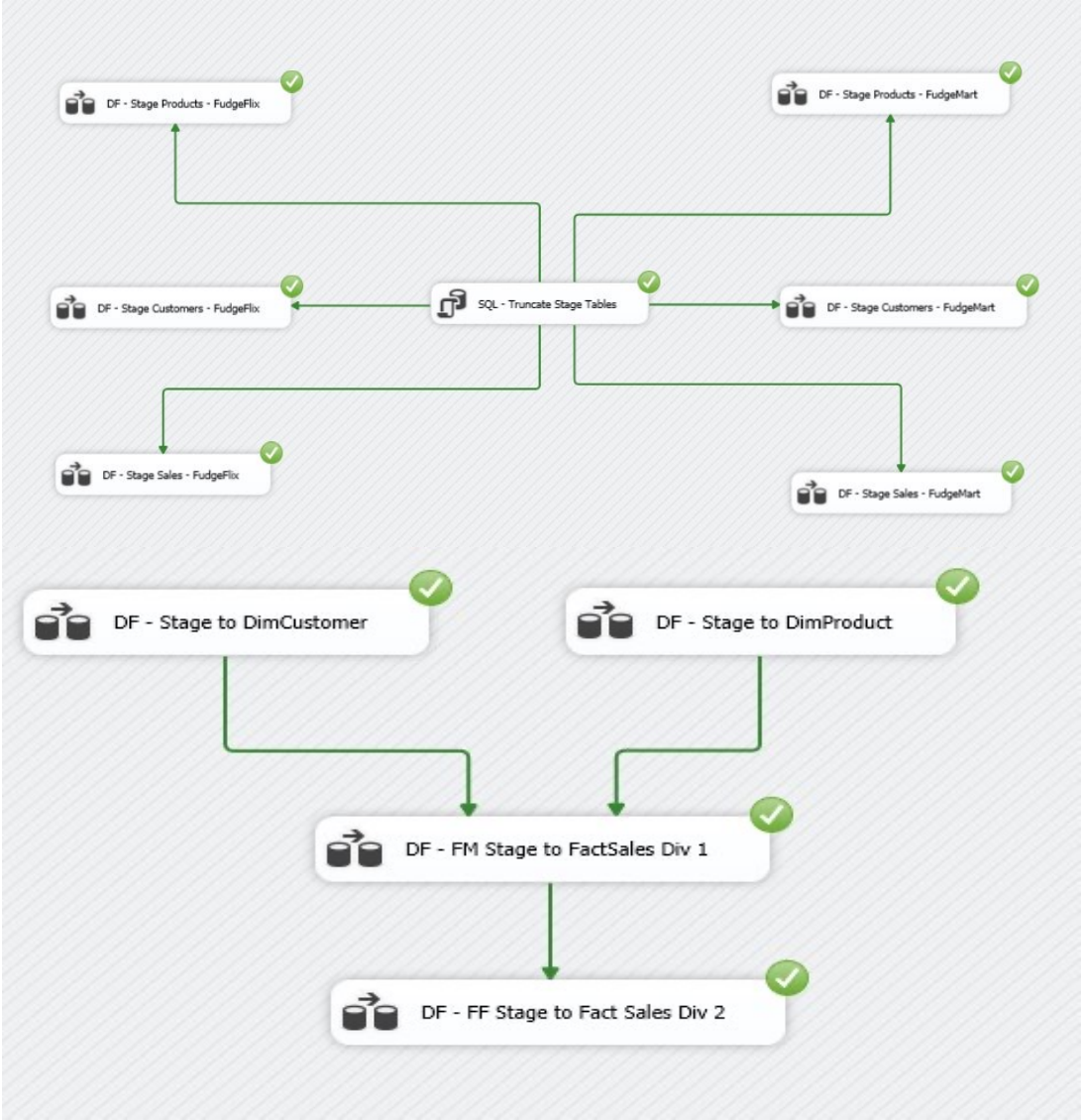
Overview

- Constructed a BI solution for the merged companies (FudgeMart and FudgeFlex)
- Built a merged data warehouse to allow data analysis from both companies and historical data
- Enabled analysis of FudgeWorld sales while drilling down into specific business line contributions



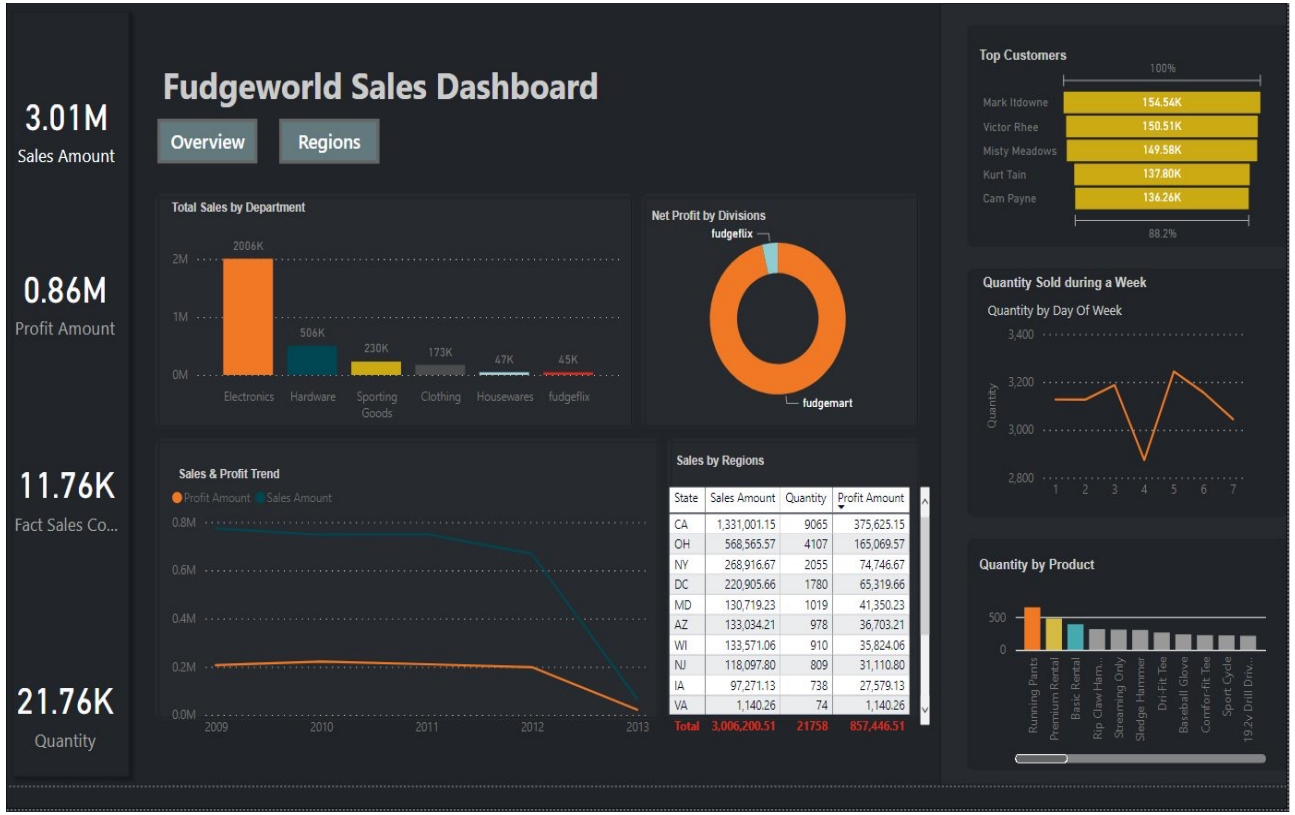
Methodology

- Constructed the FudgeWorld data warehouse using the ETL process
- Utilized SQL Server for data extraction and created a staged copy within the data warehouse
- Designed high-level dimensional modeling to outline the data structure
- Utilized SQL Server Integration Services (SSIS) to establish connections and load data tables



MOLAP and BI Dashboard construction

- Incorporated external sources (zipcodes and date variables) to create MOLAP cubes using Microsoft SSIS
- Loaded data from the staged FudgeWorld data warehouse into PowerBI
- Created dashboards to gain insights into business sales based on regions, months, and merchandise categories



Reflection of Learning Goals

- Encapsulated the program's goals: data collection, management, insight generation, and communication
- Emphasized ethical considerations and practical applications
- Developed skills in data preparation, quality assurance, and effective communication of complex insights
- Familiarized with the holistic process of data warehousing and ELT progress (Kimball and Inman models)

Project 5: Credit Card Fraud Detection

IST 707: Applied Machine Learning

Link:
https://github.com/mhgarrett/Meichan-Huang-SU-Applied-Data-Science-Portfolio-Project-Milestone-/tree/689f434cc268a996ca10f81ea5022f6f3b89d7ba/Project%205%3A%20Credit%20Card%20Fraud%20Detection

Overview

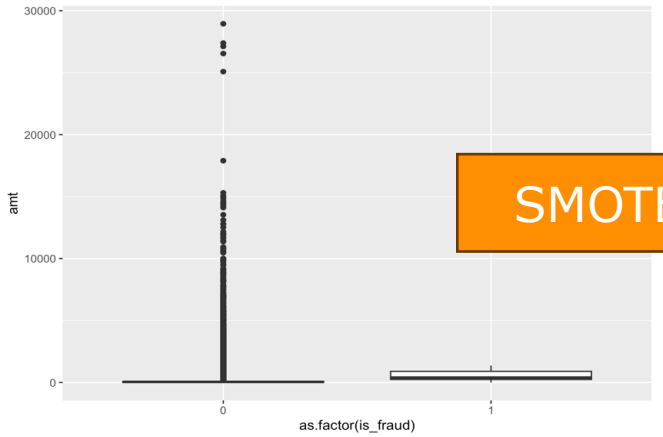
- Expanded understanding of extracting meaningful knowledge from vast dataset
- Employed different classification algorithms to predict fraud cases using a large Kaggle dataset
- Utilized R studio for data preprocessing, analysis, and machine learning algorithms

- **Dataset:**
- Simulated credit card transactions legitimate and fraud transactions from the duration 1st Jan 2019 - 31st Dec 2020 from Kaggle
- Consists of 555,719 rows of observations, which has 23 columns of variables, 12 of which are qualitative data:

```
'data.frame':   1296675 obs. of  23 variables:
 $ X                  : int  0 1 2 3 4 5 6 7 8 9 ...
 $ trans_date_trans_time: chr  "2019-01-01 00:00:18" "2019-01-01 00:00:44" "2019-01-01 00:00:51" "2019-01-01 00:01:16" ...
 $ cc_num             : num  2.70e+15 6.30e+11 3.89e+13 3.53e+15 3.76e+14 ...
 $ merchant           : chr  "fraud_Rippin, Kub and Mann" "fraud_Heller, Gutmann and Zieme" "fraud_Lind-Buckridge" "fraud_Kutch, Hermiston a
...
 $ category           : chr  "misc_net" "grocery_pos" "entertainment" "gas_transport" ...
 $ amt                : num  4.97 107.23 220.11 45 41.96 ...
 $ first              : chr  "Jennifer" "Stephanie" "Edward" "Jeremy" ...
 $ last               : chr  "Banks" "Gill" "Sanchez" "White" ...
 $ gender             : chr  "F" "F" "M" "M" ...
 $ street             : chr  "561 Perry Cove" "43039 Riley Greens Suite 393" "594 White Dale Suite 530" "9443 Cynthia Court Apt. 038" ...
 $ city               : chr  "Moravian Falls" "Orient" "Malad City" "Boulder" ...
 $ state              : chr  "NC" "WA" "ID" "MT" ...
 $ zip                : int  28654 99160 83252 59632 24433 18917 67851 22824 15665 37040 ...
 $ lat                : num  36.1 48.9 42.2 46.2 38.4 ...
 $ long               : num  -81.2 -118.2 -112.3 -112.1 -79.5 ...
 $ city_pop           : int  3495 149 4154 1939 99 2158 2691 6018 1472 151785 ...
 $ job                : chr  "Psychologist, counselling" "Special educational needs teacher" "Nature conservation officer" "Patent attorney"
 $ dob                : chr  "1988-03-09" "1978-06-21" "1962-01-19" "1967-01-12" ...
 $ trans_num          : chr  "0b242abb623afc578575680df30655b9" "1f76529f8574734946361c461b024d99" "a1a22d70485983eac12b5b88dad1cf95"
"6b849c168bdad6f867558c3793159a81" ...
 $ unix_time          : int  1325376018 1325376044 1325376051 1325376076 1325376186 1325376248 1325376282 1325376308 1325376318 1325376361 .
 $ merch_lat          : num  36 49.2 43.2 47 38.7 ...
 $ merch_long         : num  -82 -118.2 -112.2 -112.6 -78.6 ...
 $ is_fraud           : int  0 0 0 0 0 0 0 0 0 0 ...
```

Data Preparation

- Addressed data imbalance using Synthetic Minority Over-sampling Technique (SMOTE)
- Standardized categorical variables and normalized numerical data ranges
- Applied target encoding to manage high dimensionality of categorical variables
- Utilized logarithmic transformation to normalize numerical data with wide-ranging values

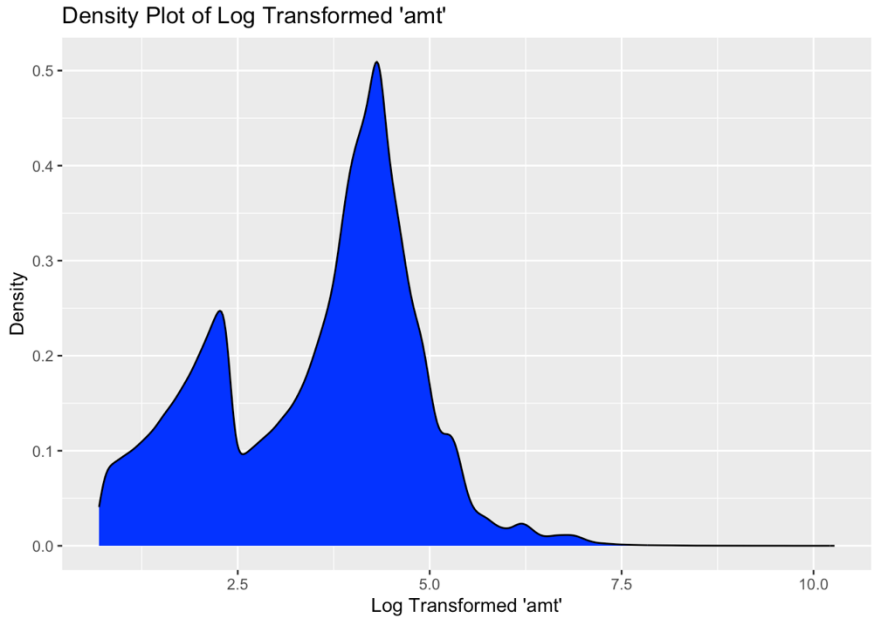


SMOTE

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

table(smote_train$is_fraud)
```

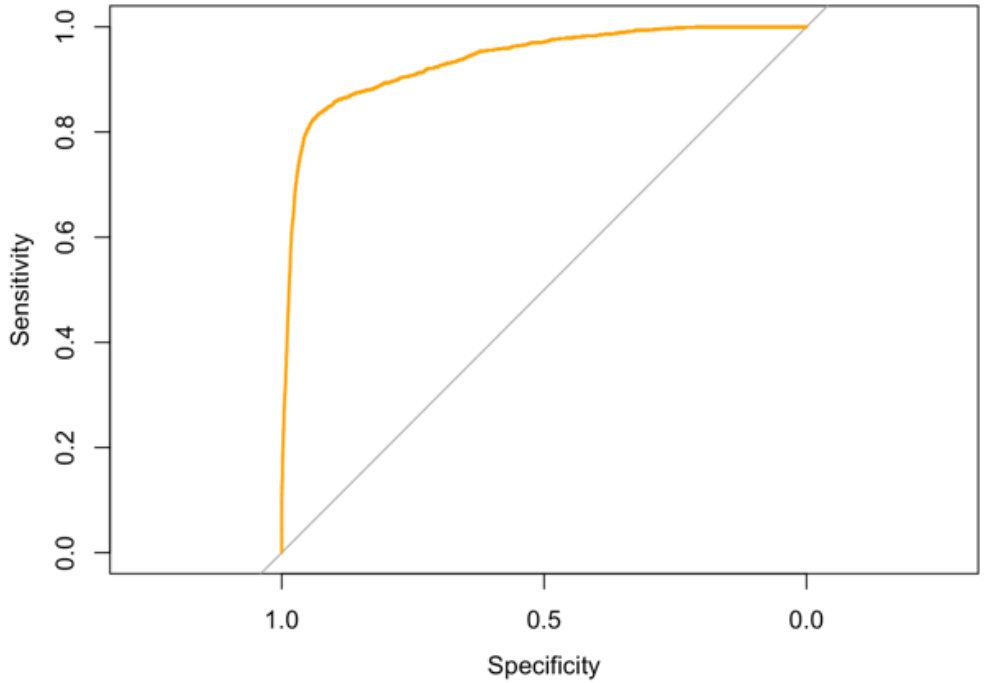
```
    0     1
30024 22518
```



Model Evaluation

Employed alternative metrics (precision, recall, F1 scores, AUC) for comprehensive model assessment

Tested six conditions: Imbalanced vs. SMOTE data using logistic regression, decision tree, and random forest

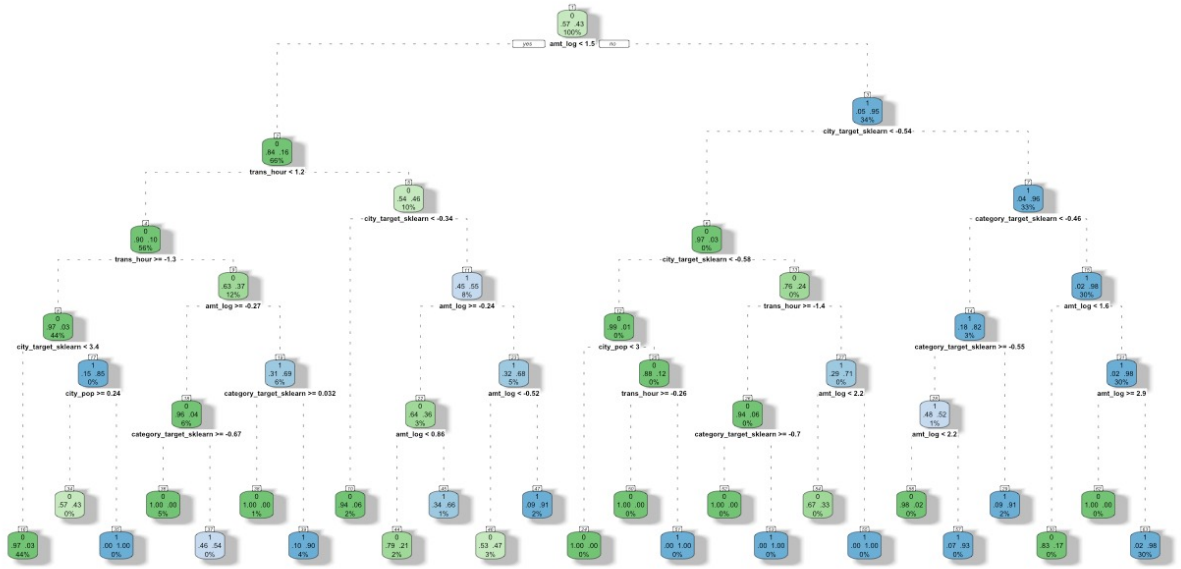


Logistic Regression	Imbalanced data	SMOTE data
Accuracy	0.975	0.887
Precision	0.097	0.0289
Recall (sensitivity)	0.65	0.88
F1 score	0.169	0.056

Table: Comparative performance matrices of imbalanced and SMOTE data using Logistic Regression

Results and Insights

- Random Forest model using SMOTE data identified as the most effective in fraud detection
- Identified attributes with strong indications of fraudulent activities (transaction categories, age, timing, location)
- Gender not found to be a strong indicator of fraud in the dataset



Rattle 2023-Dec-15 11:04:29 meichanhuang

Reflection of Learning Goals

- Solidified understanding of data science principles and their application to complex real-world problems
- Emphasized the crucial role of data preparation in the broader context of data science
- Reinforced the importance of diligently curated data for accurate analysis

Conclusion

- Transformed from a novice to a future data scientist capable of informed decision-making
- Gained extensive hands-on experience in collecting, storing, and accessing data
- Developed ability to create actionable insights applicable across diverse contexts
- Engaged with the full data science lifecycle, applying visualization techniques and predictive models
- Growing proficiency in data analytic tools, i.e. programming languages (R, Python, SQL) to manipulate data and reveal underlying stories; and business intelligence solutions (Microsoft SISS; Microsoft SQL Server; Power BI)

Syracuse University

Closing Text

Meichan Huang
PhD in Applied Linguistics/MSc in Applied Data Science
mhuang01@syr.edu; meichan.huangg@gmail.com
LinkedIn: https://www.linkedin.com/in/meichanhuang/
GitHub: https://github.com/mhgarrett