MSADS Portfolio Milestone

SYRACUSE UNIVERSITY MARCH 2024

Chinmay Ashok Maganur

SUID : 370801446

Mail : camaganu@syr.edu

Table of Contents

1. Introduction

The Applied Data Science program at Syracuse University's School of Information Studies equips students with the skills to collect, manage, analyze, and derive insights from data across various domains using diverse tools and techniques. Through courses such as Database Administration(IST 659), Introduction to Data Science (IST 687), Natural Language Processing (IST 664), and Text Mining - Mining (IST 736), students develop reports and presentations that deliver insights by leveraging Microsoft Access, SQL Server Management Studio, Python, R, NLP, and Tableau.The program's curriculum empowers data scientists specializing in marketing analytics to generate value within their organizations and provide actionable recommendations.

The Applied Data Science Program encompasses seven learning objectives, which are exemplified by the applications in this portfolio:

1. Gain a comprehensive overview of the major practice areas in data science.
2. Acquire skills to collect and organize data effectively.
3. Identify patterns in data through visualization, statistical analysis, and data mining techniques.
4. Develop alternative strategies based on data-driven insights.
5. Formulate a plan of action to implement business decisions derived from analyses.
6. Demonstrate effective communication skills regarding data and its analysis for relevant professionals within the organization.
7. Synthesize the ethical dimensions of data science practice.

2. IST 659: Database Administration: Logistic Supply Database

Video Working :

Github Link

The primary objective of this project is to develop a comprehensive database system for supply chain management, enabling suppliers to continuously monitor relevant statistics and facilitate efficient interactions with customers. The key aspects of this project include:

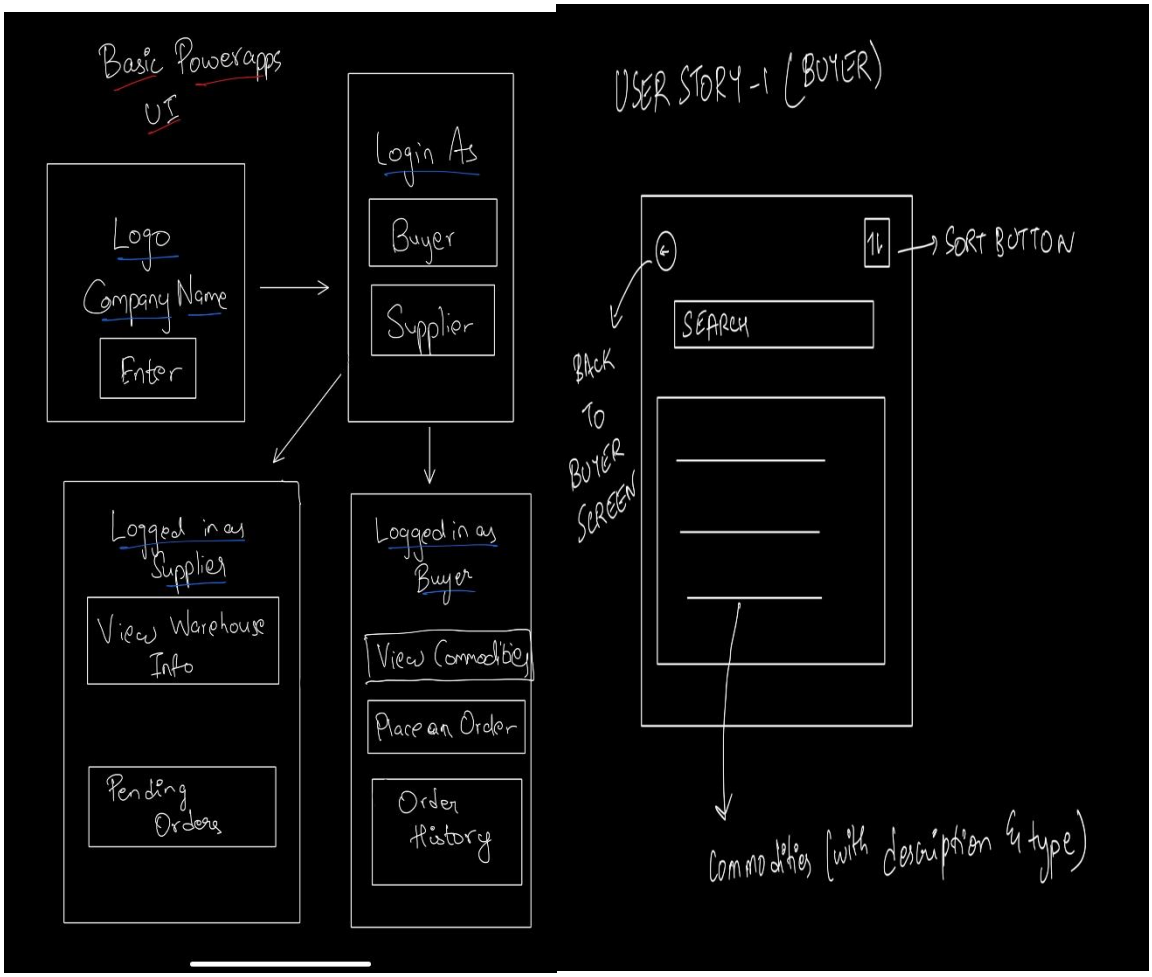
- Record Management: Implement a robust system for monitoring and managing records related to warehouses, stock levels, orders, and other critical supply chain components.
- Supplier Interaction: Develop a user-friendly interface that allows suppliers to access and analyze necessary statistics and real-time insights, empowering them to make informed decisions.
- Customer Experience: Integrate a dynamic user interface that continuously updates and provides accurate information about available commodities, enabling customers to place and modify orders efficiently.

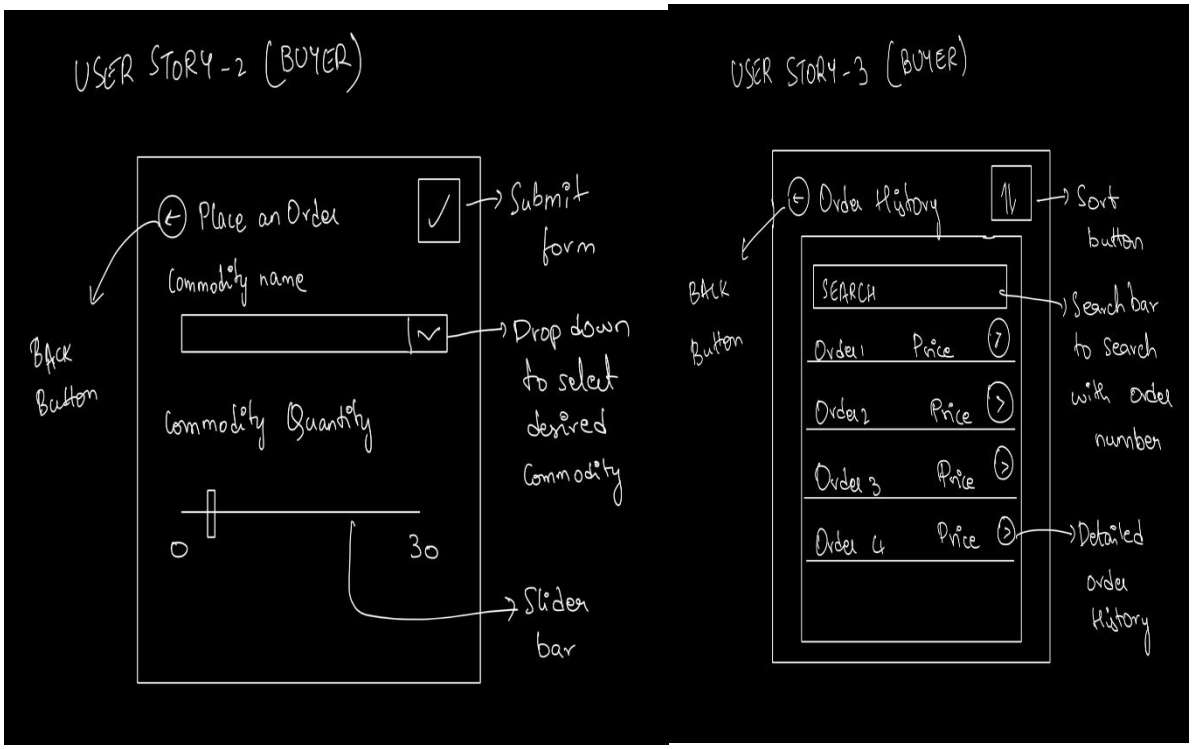
Project Specifications:

- Entity-Relationship (ER) Diagram: Develop an ER diagram to represent the data requirements and business rules for the supply chain management system, ensuring a clear understanding of the entities and their relationships.
- Conceptual Data Model: Create a conceptual data model that provides a high-level representation of the data entities, their attributes, and relationships, facilitating a shared understanding among stakeholders.
- Logical Data Model: Translate the conceptual data model into a logical data model, which will serve as a blueprint for the physical database implementation, ensuring data integrity and efficiency.
- SQL Scripts: Develop SQL scripts for creating, modifying, and dropping tables, as well as implementing keys and constraints, ensuring data consistency and integrity within the database.
- User Stories: Gather and document user stories that capture the requirements and expectations of various stakeholders, including suppliers and customers, to drive the development of the user interface and functionality.

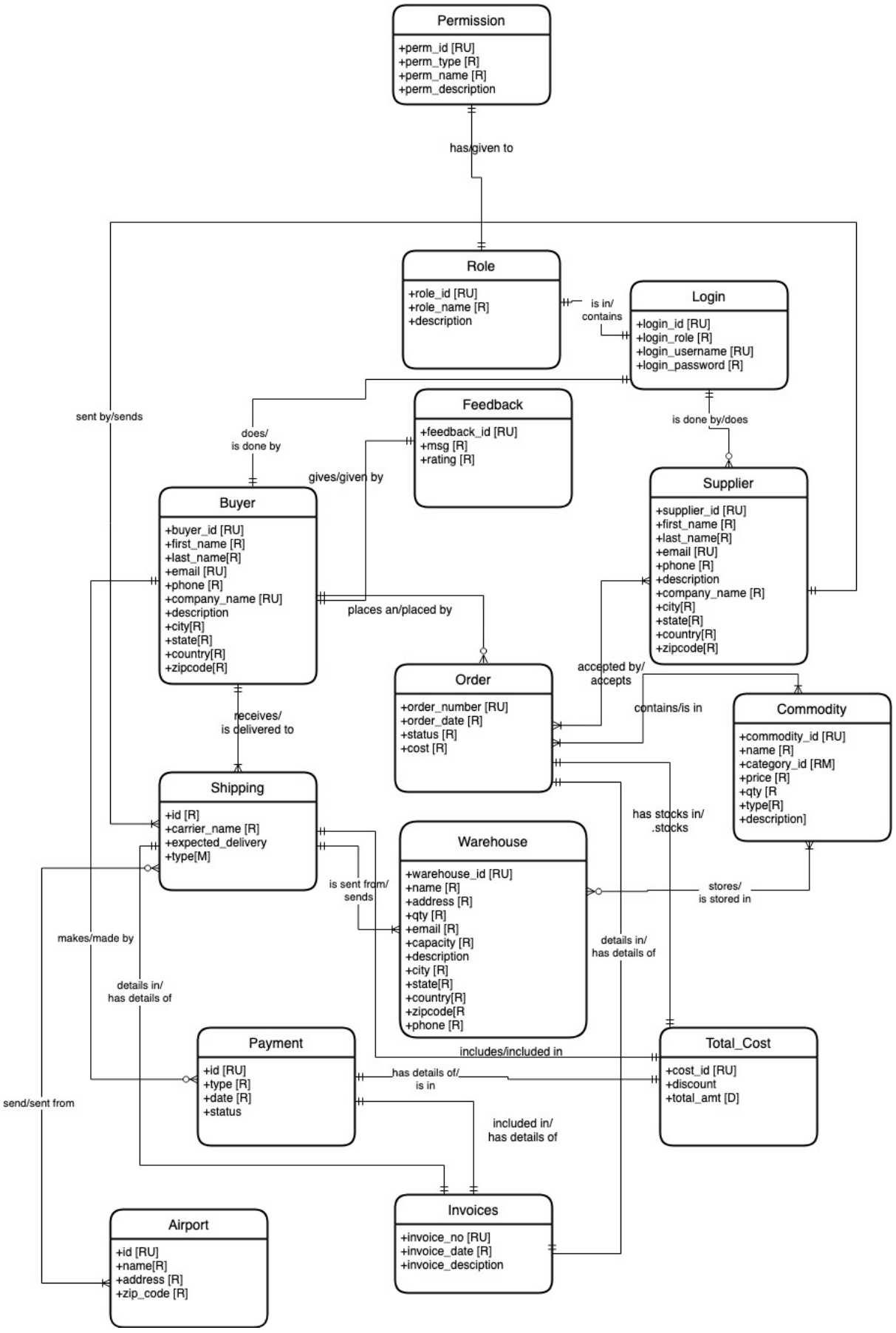
- External Data Model: Construct an external data model that aligns with the user stories, ensuring that the database design and user interface cater to the specific needs and use cases of the target users.
- User Interface Design: Develop a user-friendly and intuitive interface that enables suppliers to access and analyze relevant data, and allows customers to seamlessly place and modify orders based on real-time information.

User Stories:




USER STORY-2 (BUYER)
Place an Order
Submit form
Commodity name
Drop down to select desired Commodity
Back Button
Commodity Quantity
0
30
Slider bar
USER STORY-3 (BUYER)
Order History
Sort button
Back Button
SEARCH
Search bar to search with order number
Order 1
Price
Order 2
Price
Order 3
Price
Order 4
Price
Detailed order History

Conceptual Data Model



Reflection & Learning Goals

This supply chain management database project provides an opportunity for significant learning and growth in various aspects of data management and application development. Gained practical experience in creating conceptual and logical data models, enabling a deeper understanding of data relationships and data integrity principles.Developed proficiency in translating business requirements into well-structured database designs, ensuring data consistency and effective storage. Enhanced SQL scripting skills by writing complex queries, stored procedures, and triggers to manage and manipulate data efficiently.Understood and apply database normalization techniques to optimize data storage and eliminate redundancies.

3. IST 687 Introduction to Data Science - Heathcare Cost Prediction

Github Link

In my first semester project for the Introduction to Data Science course, I embarked on an analytical journey to uncover insights within medical data for a Health Management Organization (HMO). This endeavor not only solidified my foundational skills in data science but also highlighted its pivotal role in transforming healthcare.

Leveraging R, a powerful programming language for statistical computing and graphics, I navigated through the complexities of data analysis with proficiency. I became adept at utilizing various R libraries such as ggplot2 for data visualization, dplyr for data manipulation, and caret for applying machine learning algorithms. These tools were instrumental in preprocessing data, modeling, and deriving insights.

The project underscored the significance of data science in healthcare by analyzing patient data to forecast healthcare spending. By identifying patterns and correlations within the dataset, I contributed to developing strategies aimed at reducing healthcare costs for the HMO, showcasing the potential of data-driven decision-making in improving health management practices.

Key Findings:

- Age and smoking habits were major predictors of healthcare spending, with older individuals and smokers incurring higher costs.
- Active individuals who regularly exercise tend to have lower healthcare costs.
- Obesity, indicated by a higher BMI, was associated with increased healthcare expenses.

My application of machine learning techniques was a cornerstone of this project. I employed:

- Linear Regression to identify relationships between various factors and healthcare costs.
- Support Vector Machines (SVM) to classify patients based on predicted healthcare spending, providing a nuanced understanding of cost drivers.
- Decision Trees to visually represent decision processes, making the findings accessible to non-technical stakeholders.

These methodologies enabled me to pinpoint critical predictors of healthcare expenditure, such as age, smoking status, and BMI, thus offering actionable insights to the HMO.

Diverse Data Collection: Ensuring datasets are representative of the broad population to prevent algorithmic biases.

Transparent Methodologies: Adopting transparent and explainable AI to make the workings of algorithms understandable to users and stakeholders, facilitating trust and accountability.

Ethical Oversight: Establishing ethics boards and review processes to evaluate the ethical implications of data projects in healthcare.

Reflection & Learning Goals

This project was a profound exploration of the intersection between data science and healthcare. It not only enhanced my technical prowess in R programming and machine learning but also deepened my understanding of the ethical dimensions of data science. The insights generated have potential implications for policy-making and operational strategies in healthcare organizations, demonstrating the transformative power of data science in real-world contexts.

4. IST 707 - Applied Machine Learning - Fraud Detection in Credit Card Transactions

Github Link

In my second semester, I undertook a challenging project focused on detecting fraudulent activities in credit card transactions as part of my Machine Learning course. This project was instrumental in honing my skills in applying sophisticated machine learning techniques to combat financial fraud.

Utilizing Python and its powerful libraries, we embarked on a multifaceted approach involving exploratory data analysis (EDA), data preprocessing, feature engineering, and predictive modeling, outlined as follows:

EDA and Data Preprocessing

Initial data examination revealed a dataset with 786,363 records and 29 fields, from which irrelevant and null-containing fields were dropped. We focused on understanding the dataset's structure, exploring categorical, numerical, and boolean variables to prepare the data for modeling.

Data Wrangling

We devised algorithms to identify multi-swipe and reversed transactions, crucial for understanding patterns that could indicate fraudulent activity.

Predictive Modeling

Employed a range of machine learning models including Decision Tree, Random Forest, Logistic Regression, and XG-Boost, with a focus on handling the dataset's imbalance and optimizing for fraud detection accuracy.

Predictive Modelling for Credit Card Fraud

How does credit card fraud happen?

- Lost/Stolen card
- Card details overseen by another person
- Hacked bank details

Main challenges involved in credit card fraud detection are:

- Enormous size of Data: the model build must be fast enough to respond to the scam in time.
- Imbalanced Data: most of the transactions are not fraudulent which makes it really hard for detecting the fraudulent ones.
- Data availability as the data is mostly private.
- Misclassified Data can be another major issue, as not every fraudulent transaction is caught and reported.

How to find a way around?

- The model used must be **simple and fast** enough to detect the anomaly and classify it as a fraudulent transaction as quickly as possible.
- Use AUPRC metric to evaluate performance
- Since a synthetic dataset is readily provided, the issue of availability is solved beforehand.

Plan of Action:

- **Random Forest** and **Logistic Regression** algorithms offers the required amount of complexity, while keeping the algorithm simple and transparent without compromising on speed. Ensemble methods like XGBoost will also be tested along with a simpler decision tree, for comparison.
- The data will be pre-processed, omitting unwanted columns, adding new derived features, normalisations and encodings.
- The model will be trained, tested and evaluation metrics examined.

Comparison of metrics:

Model	Accuracy	Precision	Recall	F1-score	AUPRC
Decision Tree	0.97	0.98	0.98	0.98	0.0079
Logistic Regression	0.98	0.98	1.0	0.98	0.066
Random Forest	0.98	0.98	1.0	0.98	0.507
XGBoost	0.98	0.98	1.0	0.99	0.714

Insights and Impact

The project's outcomes were twofold: first, it demonstrated the efficacy of Random Forest and SMOTE in identifying fraudulent transactions with high accuracy. Second, it provided valuable insights into the characteristics and patterns of fraudulent activities, aiding in the development of robust fraud prevention strategies.

Throughout this project, ethical considerations were paramount. The sensitive nature of financial transaction data necessitated strict adherence to privacy and data protection principles. Furthermore, the model was designed to minimize false positives, which could lead to undue inconvenience for legitimate card users.

Reflection & Learning Goals

This project was a testament to the power of machine learning in detecting and preventing credit card fraud. It not only enhanced my technical skills in Python and machine learning but also underscored the importance of ethical considerations in handling sensitive data. The knowledge and experience gained from this project has profound implications for my future endeavors in data science, especially in financial security.

5. IST 664 Natural Language Processing - Classification of Sincere and Insincere question on Quora

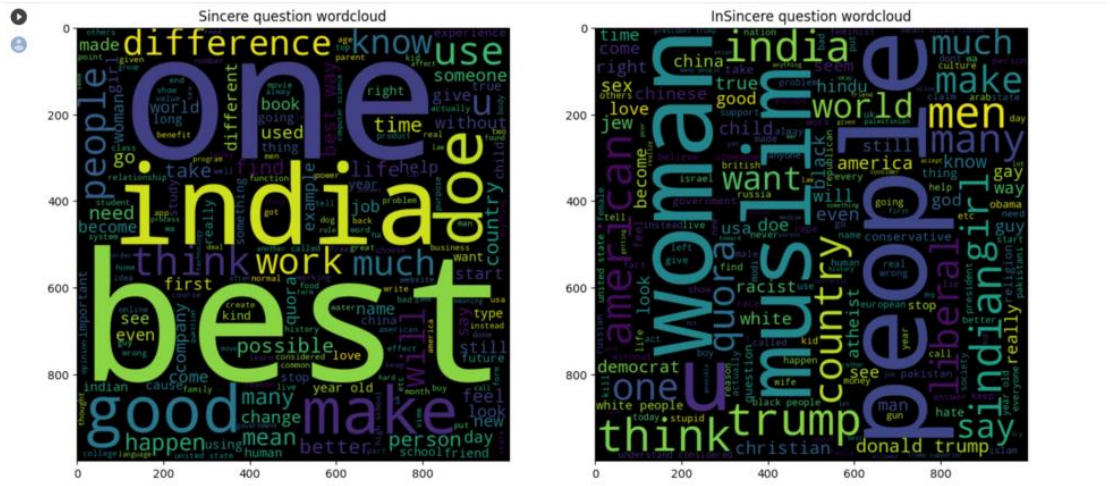
Github Link

In the second semester of my journey into Applied Data Science, I delved deep into the realm of Natural Language Processing (NLP), a field that I developed a profound interest in. My project centered on developing a sophisticated model to differentiate between sincere and insincere questions on Quora, addressing the prevalent issue of misinformation and deceitful content on online platforms.

The project was a testament to the versatility of Python in handling NLP tasks. I engaged with a plethora of NLP libraries and techniques, including:

- Data Preprocessing: Utilized pandas for data manipulation, transforming and cleaning the dataset to ensure quality input for model training.
- TF-IDF: Applied Term Frequency-Inverse Document Frequency (TF-IDF) to weigh the importance of words within the dataset, emphasizing the significance of less frequent, more informative terms.
- Machine Learning Models: Experimented with a range of models from traditional logistic regression to advanced neural networks like Bidirectional Long Short-Term Memory (Bi-LSTM).
- BERT: Leveraged the Bidirectional Encoder Representations from Transformers (BERT) model for its state-of-the-art performance in NLP tasks. This involved fine-tuning a pre-trained BERT model to suit the specific nuances of our dataset.

One of the major challenges was handling the vast and imbalanced dataset, which necessitated thoughtful preprocessing and innovative modeling strategies. Through techniques like SMOTE for oversampling and the strategic use of BERT, we were able to navigate these challenges effectively.



Model	Precision	Recall	Accuracy	F1_Score
Logistic Regression	0.66	0.37	0.94	**0.47**
Bi_LSTM	0.79	0.62	.76	**0.63**
BERT	0.81	0.82	0.93	**0.81**

Reflection & Learning Goals

The project was guided by a strong ethical framework, considering the implications of incorrectly labeling questions as insincere, which could suppress legitimate inquiries. We aimed for a balanced approach that minimized false positives while effectively filtering insincere content.

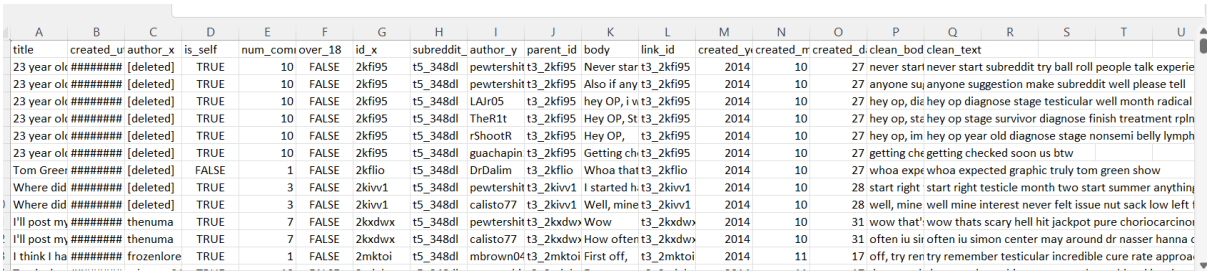
This project not only enhanced my technical skills in Python and NLP but also highlighted the potential of NLP in moderating online content and fostering healthier digital spaces. Going forward, I am keen on exploring more complex models like GPT and XLNet, and applying these NLP techniques to a broader range of applications beyond Quora, including social media sentiment analysis, customer feedback interpretation, and more.

The "Quora Sincere and Insincere Questions Filtering" project was a significant milestone in my academic and professional journey, marking my transition into an NLP enthusiast. It underscored the transformative potential of NLP in addressing real-world challenges and reinforced my commitment to pursuing excellence in this dynamic field of study.

6. IST 736 - Text Mining Insights from Cancer-Related Discussions on Reddit

[Github Link]

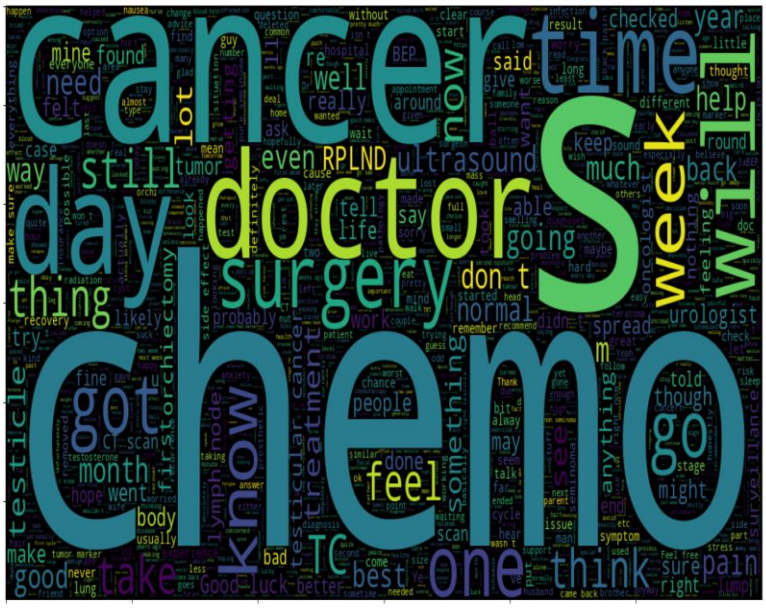
In my exploration of Text Mining during the Third semester, I engaged in a project that sought to mine insights from cancer-related discussions, specifically focusing on testicular cancer, on Reddit. This endeavor aimed to distill valuable insights from personal stories, advice, and support shared within these digital communities, offering a unique perspective on the collective experience surrounding cancer.



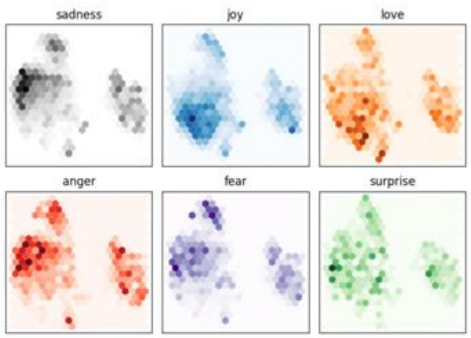
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
title	created_u	author_x	is_self	num_com	over_18	id_x	subreddit_	author_y	parent_id	body	link_id	created_y	created_m	created_d	clean_bod	clean_text				
23 year ol	########	[deleted]	TRUE	10	FALSE	2kfi95	t5_348dl	pewtershit	t3_2kfi95	Never star	t3_2kfi95	2014	10	27	never start	never start subreddit try ball roll people talk experie				
23 year ol	########	[deleted]	TRUE	10	FALSE	2kfi95	t5_348dl	pewtershit	t3_2kfi95	Also if any	t3_2kfi95	2014	10	27	anyone su	anyone suggestion make subreddit well please tell				
23 year ol	########	[deleted]	TRUE	10	FALSE	2kfi95	t5_348dl	LAJr05	t3_2kfi95	hey OP, i w	t3_2kfi95	2014	10	27	hey op, dia	hey op diagnose stage testicular well month radical				
23 year ol	########	[deleted]	TRUE	10	FALSE	2kfi95	t5_348dl	TheR1t	t3_2kfi95	Hey OP, St	t3_2kfi95	2014	10	27	hey op, sta	hey op stage survivor diagnose finish treatment rpln				
23 year ol	########	[deleted]	TRUE	10	FALSE	2kfi95	t5_348dl	rShootR	t3_2kfi95	Hey OP,	t3_2kfi95	2014	10	27	hey op, im	hey op year old diagnose stage nonsemi belly lymph				
23 year ol	########	[deleted]	TRUE	10	FALSE	2kfi95	t5_348dl	guachapin	t3_2kfi95	Getting ch	t3_2kfi95	2014	10	27	getting che	getting checked soon us btw				
Tom Greer	########	[deleted]	FALSE	1	FALSE	2kflio	t5_348dl	DrDalim	t3_2kflio	Whoa that	t3_2kflio	2014	10	27	whoa expe	whoa expected graphic truly tom green show				
Where did	########	[deleted]	TRUE	3	FALSE	2kivv1	t5_348dl	pewtershit	t3_2kivv1	I started h	t3_2kivv1	2014	10	28	start right	start right testicle month two start summer anythin				
Where did	########	[deleted]	TRUE	3	FALSE	2kivv1	t5_348dl	calisto77	t3_2kivv1	Well, mine	t3_2kivv1	2014	10	28	well, mine	well mine interest never felt issue nut sack low left				
I'll post my	########	thenuma	TRUE	7	FALSE	2kxdwx	t5_348dl	pewtershit	t3_2kxdwx	Wow	t3_2kxdwx	2014	10	31	wow that'	wow thats scary hell hit jackpot pure choriocarcino				
I'll post my	########	thenuma	TRUE	7	FALSE	2kxdwx	t5_348dl	calisto77	t3_2kxdwx	How often	t3_2kxdwx	2014	10	31	often iu sir	often iu simon center may around dr nasser hanna				
I think I ha	########	frozenlore	TRUE	1	FALSE	2mktoi	t5_348dl	mbrown04	t3_2mktoi	First off,	t3_2mktoi	2014	11	17	off, try ren	try remember testicular incredible cure rate approa				

This project harnessed Python for text mining, employing a range of libraries and techniques to process and analyze the data:

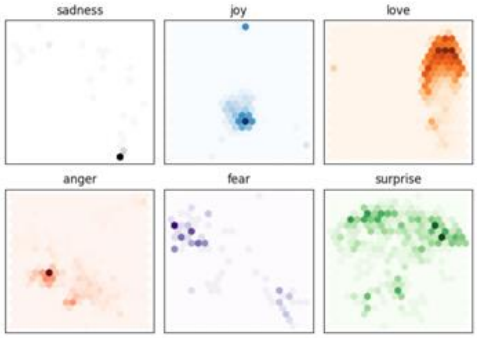
- Data Preprocessing: Implemented comprehensive data cleaning techniques including stop word removal, lemmatization, and tokenization to prepare the data for analysis.
- Emotion Classification: Developed a classifier to categorize text by emotion, utilizing algorithms like Logistic Regression and Gaussian Naive Bayes, and enhancing model performance through fine-tuning.
- Topic Modeling: Applied advanced methods like TF-IDF with LDA, BERTopic, and Kmeans clustering combined with TF-IDF to uncover underlying themes in the discussions.



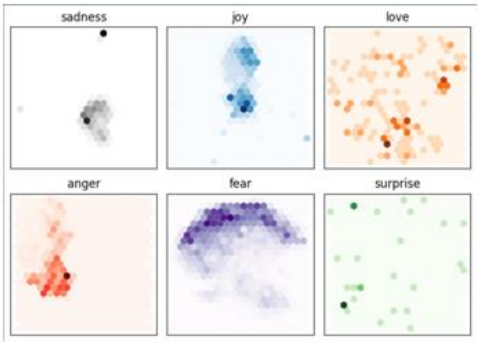
7



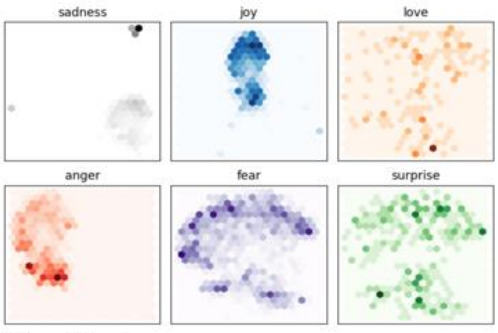
Original Data



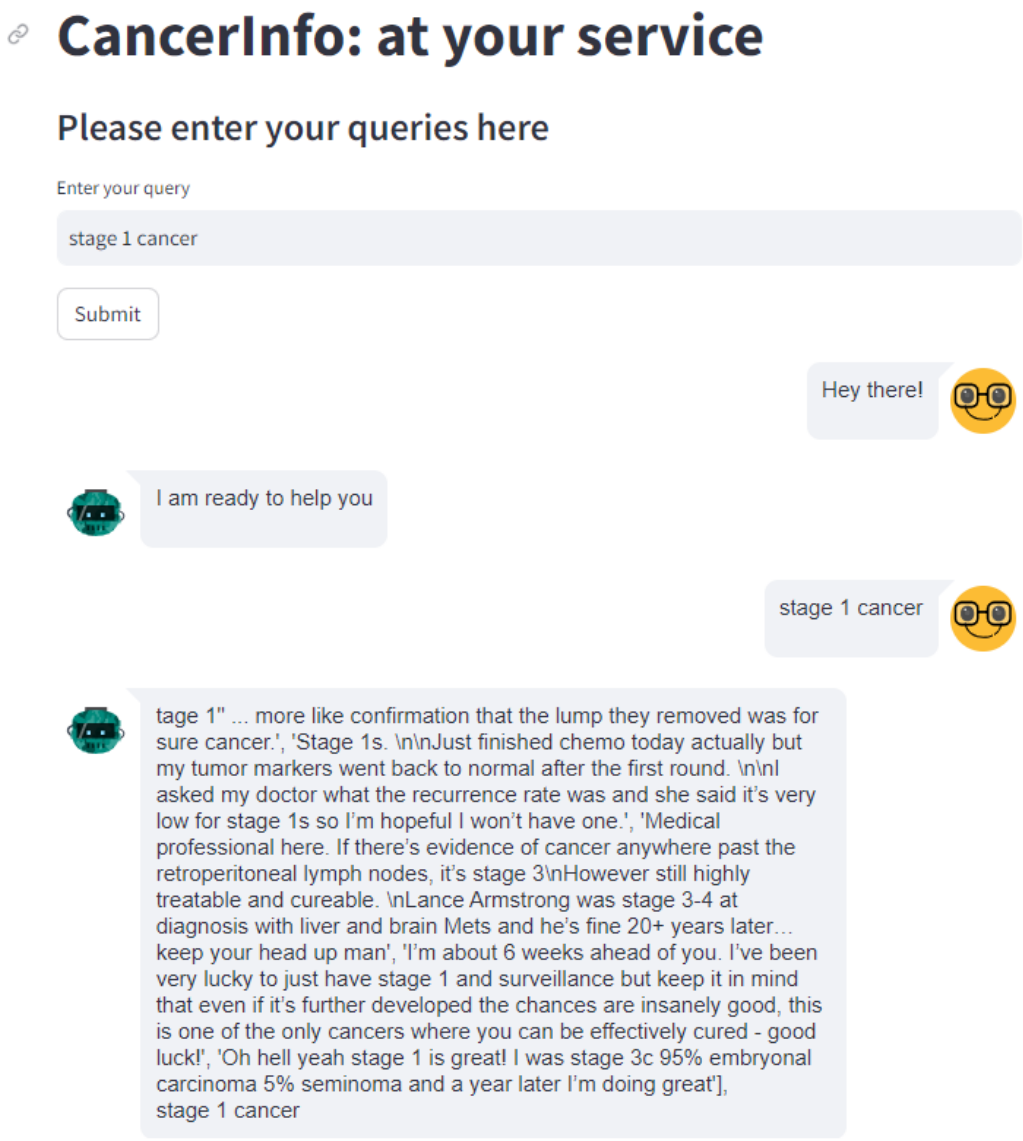
Logistic Regression



G MNB



Fine Tuning



Challenges and Solutions

Navigating the complex nature of unstructured text data presented significant challenges, particularly in accurately classifying emotions and identifying coherent topics. Through iterative model refinement and the incorporation of advanced NLP techniques, these obstacles were addressed, leading to meaningful categorization and insights.

Ethical Considerations

Throughout the project, ethical considerations were paramount. Ensuring privacy and consent, the project utilized publicly available data, anonymized to protect individual identities. Efforts were made to mitigate biases in the models and algorithms, maintaining the integrity and objectivity of the analysis.

Findings and Implications

- The project revealed distinct emotional patterns and topics within the cancer-related discussions:
- Emotion Analysis: The classifier identified a predominant presence of joy, followed by sadness and love, offering insights into the emotional landscape of the discussions.
- Topic Insights: Topic modeling unveiled key themes such as health concerns, post-surgical experiences, diagnosis and treatment discussions, and emotional support, highlighting the multifaceted nature of the conversations.

This project illuminated the power of text mining in understanding complex human experiences shared online. By analyzing cancer-related discussions on Reddit, we gained insights into the prevalent emotions and topics, contributing to a deeper understanding of the collective narratives surrounding cancer. These findings have the potential to inform healthcare providers, support groups, and policymakers, ultimately enhancing patient care and support systems.

7. Conclusion:

Throughout the Applied Data Science program at Syracuse University, I have embarked on a transformative educational journey, cultivating a robust skill set that spans the full spectrum of data science applications. The diversity of projects undertaken has solidified my technical expertise while emphasizing the practical implications of data science across various domains.

The projects presented in this portfolio exemplify the comprehensive learning objectives of the program, showcasing my ability to collect and organize data, identify patterns through sophisticated analytical techniques, and derive actionable insights that influence strategic decision-making. From developing database systems for enhancing supplier-customer interactions to predicting healthcare costs and identifying key cost drivers, these projects have refined my technical skills in using tools like R, Python, and various machine learning algorithms.

Moreover, the application of machine learning to detect fraudulent transactions and the use of natural language processing to filter sincere from insincere content on digital platforms illustrate the versatility and impact of data-driven methodologies. Each project has been a stepping stone towards understanding the ethical dimensions of data science, ensuring the responsible use of technology and data in solving real-world problems. As I move forward, I am committed to leveraging this expertise to foster advancements in technology and business, ensuring that data science continues to be a force for positive change in society.

8 . References

- https://medium.com/analytics-vidhya/credit-card-fraud-detection-in-python-using-scikit-learn-f9046a030f50
- https://github.com/jbofill10/C1_Transaction_Data
- https://dsproject-ist687-group3.shinyapps.io/ShinyApp_Group3_IST687/
- https://towardsdatascience.com/fine-tuning-bert-for-text-classification-54e7df642894#6ba6
- https://www.analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/?utm_source=blog&utm_medium=comprehensive-guide-attention-mechanism-deep-learning
- https://towardsdatascience.com/multi-label-text-classification-using-bert-and-tensorflow-d2e88d8f488d
- https://arxiv.org/pdf/1810.04805.pdf
- https://pytorch.org/hub/huggingface_pytorch-transformers/
- https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270