IST 782 Applied Data Science Portfolio

By

Saikumarreddy Pochireddygari
SUID: 367190390
Email: spochire@syr.edu

Table of Contents:

1. **IST 718 – Big Data Analytics**

In my Big Data Analytics course (IST 718), I worked on a project that aimed to improve how a company named Xente, based in Uganda, spots and stops fraud. My goal was to use big data and machine learning to pick out and prevent fake transactions, making Xente safer for its users.

Project Overview: I built a system that uses big data to find fraud in financial dealings. By looking closely at the data from Xente's transactions, my aim was to make their fraud-finding methods better, keeping their services safe and trusted.

Methodology: I followed a step-by-step approach:

- **Data Handling:** I got data from Xente, cleaned it up, and got it ready for analysis.
- **Analysis:** I did a deep dive into the data to spot patterns and learn more about the transactions.
- **Feature Engineering:** I created and improved certain data features, like how often transactions happen and their sizes, to help spot fraud.
- **Model Development:** I used different computer models, like Decision Trees and Random Forest, adjusting them to catch fraud better. I paid special attention to the F1 score, a way to measure how good the models are, especially since the data wasn't balanced.

Conclusion: My project made a big difference in how well Xente can find fraud. Even though no one model was perfect for every situation, some, like GBDT and Random Forest, worked really well in certain cases. This doesn't just make Xente safer; it also makes their customers trust them more.

Reflection and Learning Goals: This project taught me a lot about analyzing data and creating models to predict outcomes. I learned how to handle challenges, like data that's not balanced, making sure my models are fair and effective. This experience is a great foundation for my future work in this field, showing how important it is to understand data and use it to solve real problems.

Linking Learning Goals to Deliverables:

- **Data Analysis Skills:** By analyzing transaction data, I demonstrated my ability to work with large datasets, a key skill in big data analytics.
- **Problem-Solving:** Developing models to detect fraud helped me show my problem-solving skills, crucial for any data scientist.
- **Technical Proficiency:** Using machine learning models showcased my technical skills, preparing me for advanced work in this area.

Preparation for Specialty Area: The project has prepared me for a career in data science, especially in areas where security is key. The skills I've developed in this course will help me in jobs where I need to analyze data and make smart decisions based on that analysis.

This project was not just an assignment; it was a chance to apply what I've learned in class to a real-world problem, enhancing my skills and preparing me for my future career.

2. IST 707 – Applied Machine Learning

In my course IST707: Applied Machine Learning, I tackled a project on sentiment analysis using Twitter data to understand public sentiment about vaccines. This was particularly aimed at analyzing sentiments related to the COVID-19 vaccine amidst various views expressed on social media.

ProjectOverview:
I conducted sentiment analysis to identify public opinions on vaccines from Twitter data. By examining tweets, I intended to help public health organizations understand public sentiment, addressing misconceptions and improving communication strategies.

Methodology:
My approach was comprehensive:

- **Data Handling:** I cleaned and prepared two datasets: one from ZINDI with 10,001 tweets and another provided by Professor King with 20,000 tweets, focusing on COVID vaccine discussions.
- **Data Analysis:** I explored the data, understanding tweet sentiments that were categorized as negative, neutral, or positive.
- **Feature Engineering:** I refined the data, removing irrelevant text and creating features like text length and word count to assist in analysis.
- **Model Development:** I applied various machine learning models like kNN and Decision Trees, alongside advanced methods like BERT Vanilla, to classify tweet sentiments, using metrics like accuracy and F1 score to evaluate model performance.

Conclusion:
My analysis provided insights into public sentiment on vaccines, showcasing the power of machine learning in processing and categorizing large amounts of social media data. The SVM model showed notable improvements in classifying sentiments, and the BERT Vanilla model, a deep learning approach, offered promising results, outperforming other models in accuracy.

Reflection:
This project was a deep dive into applied machine learning, enhancing my skills in data cleaning, feature engineering, and model selection. It underscored the importance of adapting machine learning techniques to real-world data, like the varied and informal language found on Twitter. I learned to navigate the challenges of social media data analysis, developing a nuanced understanding of sentiment analysis in public health contexts.

Preparation for Specialty area:
This project has been pivotal in preparing me for a career in data science, especially in areas intersecting public health and technology. The skills gained here will be instrumental in tackling future challenges, where understanding public sentiment can guide effective communication and policy-making.

- **Challenges:** The informal and dynamic nature of Twitter language presented unique challenges in data cleaning and model training, which were crucial learning opportunities.

Through this project, I've not only advanced my technical skills but also developed a keen understanding of how machine learning can be leveraged to inform public health strategies, a testament to the practical impact of data science in societal contexts.

3. **IST 687 – Intro to Data Science**

In my IST-687 class, I did a project where I looked at healthcare costs using data from a Health Management Organization (HMO). My goal was to figure out which customers would cost more in terms of healthcare and why that might be.

Project Summary:
I analyzed healthcare cost data to spot trends and give useful advice to a healthcare company. I checked many things like how old people are, their BMI, if they smoke, and how much they exercise to see how these factors affect healthcare costs.

Steps I followed:

1. **Getting the Data Ready:** I cleaned up the data by fixing or removing wrong or missing information. This made sure my analysis would be accurate.
2. **Looking at the Data:** I used different ways to look at the data to understand the main trends and how different things like age or smoking status are connected to healthcare costs.
3. **Building Models:** I used some math and computer models to predict who might have high healthcare costs. I tried out several methods to see which one worked best.

Conclusion Gained: I learned that whether someone smokes, how much they exercise, and their level of education can impact how much they might spend on healthcare. Smokers and people who don't exercise much tend to have higher healthcare costs.

What I Learned From it: This project helped me get better at preparing data, analyzing it, and using different models to find answers. I learned a lot about how to turn complex data into clear insights.

Skills I Showed:

1. **Analyzing Data:** I demonstrated how to sort through big datasets, find important patterns, and spot any unusual data points.
2. **Making Predictions:** I used different predictive models to estimate healthcare costs and identify what factors make some people's healthcare more expensive.
3. **Creating Insights:** I translated my data findings into practical suggestions that could help manage healthcare costs better.

Getting Ready for My Future: This project was a great fit for my interest in healthcare data and gave me hands-on experience in tackling real-world problems in healthcare costs.

- **Challenges:** Dealing with incomplete or incorrect data was tough, but I learned how to address these issues effectively.

In conclusion, this project was a fantastic learning experience, improving my skills in data analysis and giving me insights into healthcare costs, which will be very useful for my future work in this field