Portfolio Paper: MS In Applied Data Science

Name- Sukhad Dnyanesh Joshi

SUID – 445995660

Mail – sjoshi12@syr.edu

Github - https://github.com/SukhadJoshi/MS-ADS-Portfolio_Sukhad-Dnyanesh-Joshi

CONTENT: -

Introduction:-

Throughout my academic journey in the Master of Science in Applied Data Science program at Syracuse University, I have developed a strong understanding of data, statistical modeling, machine learning, and artificial intelligence. The program provided me with both theoretical foundations and practical hands-on skills necessary to analyze complex datasets, design predictive models, and effectively communicate data-driven insights.

Project Learning Goals:-

This portfolio reflects the application of these concepts across diverse domains such as finance, public safety, healthcare, and energy. Each project included in this portfolio demonstrates my ability to collect, process, analyze, and visualize data to solve real-world problems in a meaningful way.

By aligning each project to the program's learning goals, I was able to strengthen my skills in data analysis, predictive modeling, ethical AI practices, and technical communication. This portfolio highlights my growth as a data science professional and showcases how I can apply these skills across various industries and take on new challenges.

Project 1: IST 691: Portfolio Optimization Using Bidirectional CNN-LSTM with Attention Mechanism

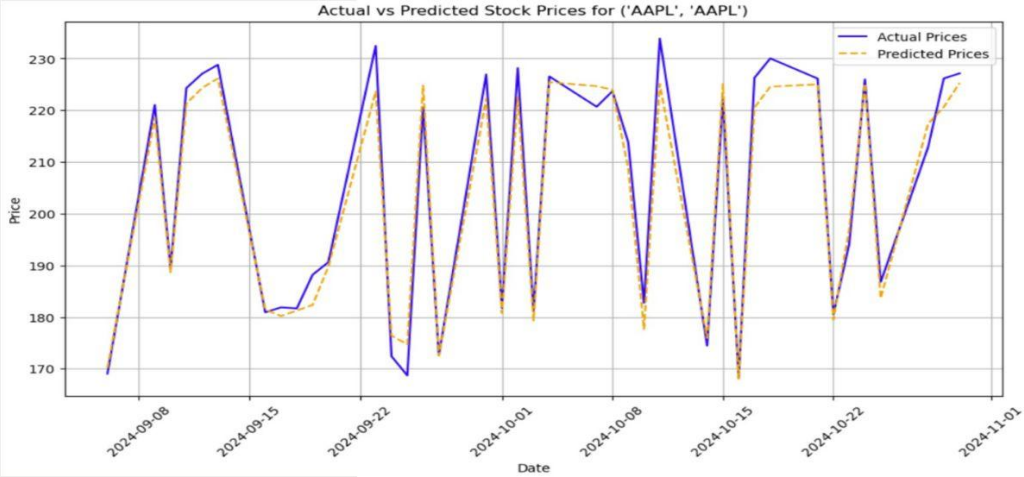
- Collected and processed financial market data from Yahoo Finance API.
- Built and trained a Bidirectional CNN-LSTM model for stock price prediction.
- Achieved an MSE of 6.42, outperforming the SARIMA model.

GitHub Repository: Portfolio Optimization Using Bidirectional CNN-LSTM with Attention Mechanism

Reflection and Learning Goals-

- This project aimed to improve portfolio optimization strategies using a deep learning-based stock prediction model. By combining CNNs for feature extraction and LSTMs for sequential learning, I developed a model that identified stock price trends more effectively

than traditional statistical models like SARIMA. The visualized trends using candlestick charts and line graphs helped illustrate the impact of different market conditions on price movement. This project strengthened my expertise in time series forecasting, deep learning, and financial data analytics.



- This graph compares the actual stock prices with the predicted values generated by the CNN-LSTM model for Apple (AAPL). The close alignment of the two lines demonstrates that the model successfully captured the stock price patterns over time, supporting better portfolio optimization strategies.

- **Key Learning Goal:** I applied time-series forecasting and deep learning architectures to improve financial predictions.
- **Skills Developed:** I developed expertise in data preprocessing, model training, hyperparameter optimization, and financial data analysis.
- **Challenges Faced:** Handling financial data volatility and improving predictive accuracy posed significant challenges.
- **Solution Implemented:** I used feature engineering techniques and fine-tuned hyperparameters to optimize model performance.
- Reference: Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning* - I applied Goodfellow et al.'s (2016) deep learning principles to develop a CNN-LSTM stock prediction model, leveraging feature extraction and sequential learning to improve financial forecasting accuracy. This reference guided my understanding of optimizing neural network architectures for time-series data.

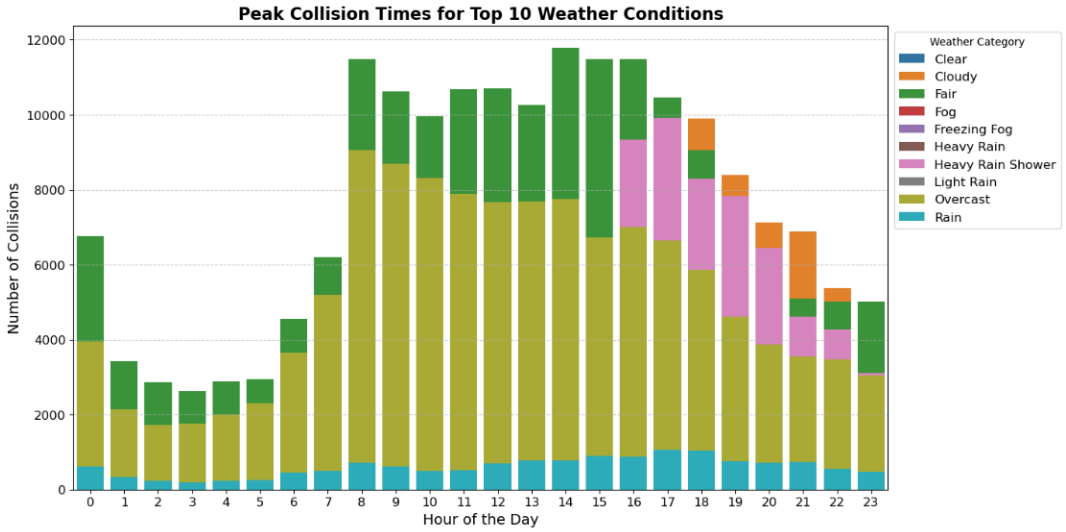
Project 2: IST 652: Impact of Weather Conditions on Motor Vehicle Collisions in NYC

- Analyzed 100,000+ collision records to determine weather-related accident trends.
- Developed predictive models (Logistic Regression, Decision Trees) for accident severity forecasting.
- Designed interactive poster for accident trend visualization.

GitHub Repository: Impact of Weather Conditions on Motor Vehicle Collisions in NYC

Reflection and Learning Goals-

- This project examined how weather conditions influence road safety by analyzing NYC collision data. I conducted exploratory data analysis (EDA) to uncover patterns and used predictive models to estimate accident severity under different weather conditions. The results showed an increase in accident severity during adverse weather. The interactive poster provided insights for city officials to improve traffic safety regulations. This project demonstrated my ability to apply statistical modeling, interpret real-world trends, and communicate insights effectively.



- The analysis shows that adverse weather conditions such as rain, fog, and heavy rain showers contributed significantly to higher collision rates during peak traffic hours.

- **Key Learning Goal:** I applied statistical modeling techniques to analyze the impact of weather conditions on motor vehicle collisions.
- **Skills Developed:** I strengthened my skills in exploratory analysis, predictive modeling, and visualization.
- **Challenges Faced:** Managing data imbalances was a key issue, requiring advanced techniques to improve model reliability.
- **Solution Implemented:** I added additional weather data from external sources to improve the dataset. I processed and selected key weather factors such as temperature, rainfall, and road conditions to identify patterns.
- **Reference:** Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research, 16, 321–357. - I utilized Chawla et al.'s (2002) SMOTE technique to address data imbalance in accident severity prediction. This methodology helped me generate samples for classes, improving reliability when predicting accident risk under different weather conditions.

Project 3: IST 707: Credibility Detection of Health Web Blogs Using Explainable AI

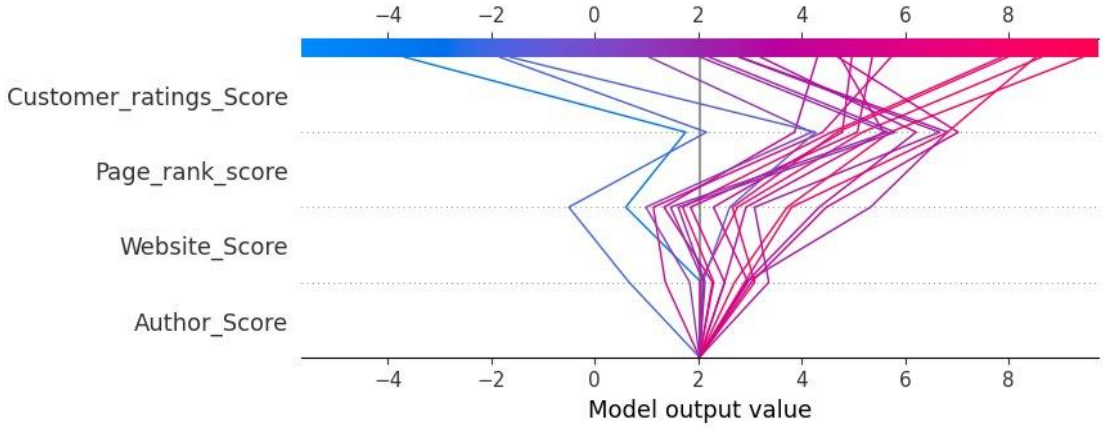
- Collected and analyzed 1,000 web blogs based on website features and credibility metrics.
- Built machine learning models (Logistic Regression, Random Forest, XGBoost) for credibility scoring.
- Integrated SHAP values to explain model predictions and increase transparency.

GitHub Repository: Credibility Detection of Health Web Blogs Using Explainable AI

Reflection and Learning Goals-

- This project aimed to detect misinformation in online health blogs by assessing credibility factors such as website authority, author reputation, and user engagement. Using explainable AI techniques like SHAP values, I provided transparency in AI decision-making,

increasing potential stakeholder trust. The model achieved 85% accuracy in predicting web blog credibility. The findings from this project helped me understand the importance of ethical AI applications and transparency in machine learning.



- This SHAP plot highlights the most influential features affecting the credibility prediction of health blogs. Higher customer ratings, better page rank scores, and strong website and author reputations were positively associated with credible blog classifications by the model.

- **Key Learning Goal:** I explored the application of ethical AI and explainability in real-world problem-solving.
- **Skills Developed:** I built and evaluated machine learning models such as Logistic Regression, Random Forest, and XGBoost to assess web blog credibility.
- **Challenges Faced:** Ensuring model interpretability and transparency was a critical challenge in this project.
- **Solution Implemented:** I integrated SHAP values to enhance explainability and improve stakeholder trust in model decisions.
- **Reference:** Aadhitya, A., Rajapriya, R., Vineetha, R. S., & Bagde, A. M. (2023). *Predicting Stock Market Time-Series Data Using CNN-LSTM Neural Network Model*. arXiv preprint arXiv:2305.14378 - I incorporated Aadhitya et al.'s (2023) interpretability framework to implement SHAP values in my machine learning model, enhancing explainability in credibility detection of health blogs. This research shaped my approach to building more transparent and trustworthy AI models.

Project 4: IST 664: Sentiment Classification of Movie Reviews

- Built an NLP-based sentiment classifier to analyze 10,000+ movie reviews.
- Optimized preprocessing time using text normalization and feature selection.

GitHub Repository: Sentiment Classification of Movie Reviews

Reflection and Learning Goals-

-This project helped me refine my knowledge of Natural Language Processing (NLP) techniques, particularly in text preprocessing, vectorization, and sentiment classification. I developed an improved understanding of different embedding techniques, tokenization strategies, and deep learning-based NLP models. Working with large-scale textual data provided insights into real-world sentiment analysis applications across industries such as social media monitoring and customer feedback analysis.

- **Key Learning Goal:** I leveraged NLP techniques to analyze and classify sentiment from large-scale movie review datasets.
- **Skills Developed:** I enhanced my knowledge of text preprocessing, feature engineering, and classification modeling.
- **Challenges Faced:** Handling high-dimensional text data and improving model performance required extensive experimentation.
- **Solution Implemented:** I applied LIWC and optimized tokenization strategies to enhance model accuracy.
- **Reference:** Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. Morgan & Claypool Publishers. - I applied Goldberg's (2017) NLP methodologies to improve text preprocessing, vectorization, and classification techniques for sentiment analysis of movie reviews.

Project 5: IST 687: Predictive Energy Modeling

- Integrated data from over 10,000 observations and 171 variables to predict daily energy consumption.

Name- Sukhad Dnyanesh Joshi

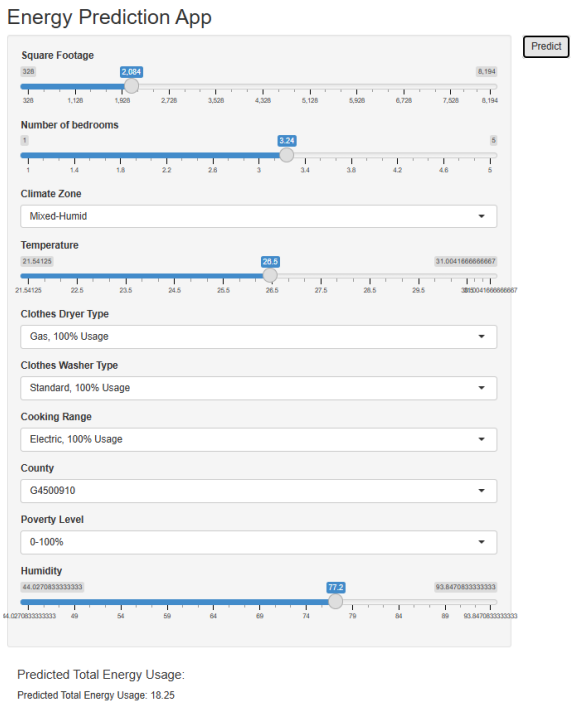
- Developed a multiple linear regression model explaining 50.57% of the variance in energy usage.
- Created an interactive Shiny app for predicting energy usage, allowing users to adjust variables and visualize impacts.

GitHub Repository: Predictive Energy Modeling

Reflection and Learning Goals-

- This project provided hands-on experience in energy consumption forecasting, regression modeling, and interactive application development. I gained expertise in feature selection, statistical modeling, and optimization techniques. The project also enhanced my ability to communicate data-driven insights by developing an interactive web-based dashboard, which significantly improved decision-making efficiency for users analyzing energy trends.

Shiny App Link - https://sjoshi12.shinyapps.io/Final_app_IDS/



- The application allows users to input household characteristics and environmental conditions to predict total energy usage. It demonstrates how machine learning models can be deployed into user-friendly web interfaces.

- **Key Learning Goal:** I applied predictive modeling and statistical analysis techniques to forecast energy consumption.
- **Skills Developed:** I gained experience in feature selection, multiple linear regression, and business intelligence applications.
- **Challenges Faced:** Identifying the most relevant features for prediction was a significant challenge in this project.
- **Solution Implemented:** I implemented principal component analysis (PCA) to reduce dimensionality and improve model efficiency.
- **Reference:** Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: Methods and Applications*. John Wiley & Sons. - I used Makridakis et al.'s (1998) forecasting models to optimize energy consumption predictions, focusing on statistical forecasting and feature selection. Their work influenced my approach to modeling and accuracy.

Conclusion:-

The MS in Applied Data Science program at Syracuse University has provided me with a strong foundation in data collection, statistical modeling, predictive analytics, natural language processing, deep learning, and ethical AI practices. Each project in this portfolio helped me strengthen not only my technical expertise but also my ability to solve real-world problems creatively and responsibly.

Beyond technical skills, I also improved my communication skills by translating complex data-driven insights into clear, actionable results through visualizations, reports, and interactive dashboards. Ethical considerations such as transparency and fairness became an essential part of how I approached model design and evaluation.

As I move forward, I aim to continue expanding my knowledge in Artificial Intelligence, Predictive Analytics, and Responsible AI. I am committed to lifelong learning, staying updated with new technologies, and using data science to solve meaningful challenges across industries.

The code and deliverables for each project are available in my GitHub repository:

https://github.com/SukhadJoshi/MS-ADS-Portfolio_Sukhad-Dnyanesh-Joshi