Syracuse University, School of Information Studies
M.S. Applied Data Science

Portfolio Milestone

Student: Benjamin J Heindl
SUID: 359536618
GitHub: Project Portfolio

Benjamin J. Heindl - 3595366618
M.S. Applied Data Science

Table of Contents

Benjamin J. Heindl - 3595366618
M.S. Applied Data Science

Introduction

In the Applied Data Science program at Syracuse University's School of Information Studies, I embarked on a comprehensive journey to harness the power of data across various domains. This program equipped me with essential skills to collect, manage, analyze, and derive insights using diverse tools and techniques. My portfolio showcases projects from key courses like Database Administration, Scripting for Data Analysis, Applied Machine Learning, Natural Language Processing and Deep Learning. These projects, developed using tools such as Microsoft Access, SQL Server Management Studio, Python, R, Excel, and Tableau, reflect my ability to deliver insightful reports and presentations.

The program's learning objectives, which include areas like data collection, pattern identification through visualization and statistical analysis, strategy development based on data, and ethical considerations in data science, are exemplified in my work. These projects not only demonstrate my technical proficiency but also my capability to produce actionable recommendations and generate value in marketing analytics. My portfolio is a testament to the comprehensive education provided by the School of Information Studies, preparing me as a data scientist capable of making informed, impactful decisions in any organization.

IST 659 - Database Administration Concepts & Database Management

SQL Tee Mates: Developing a PGA Tour Analysis Database Application

The "SQL Tee Mates" project, part of the IST 659 course, involved creating a database application for analyzing PGA tour results from 2018 to 2022. The project focused on importing data, executing SQL scripts for efficient database management, and developing a user-friendly interface for querying PGA tour results.

Key Aspects

Data Importation: The process included importing CSV files into Azure Data Studio to ensure efficient database functionality.

SQL Scripting: Emphasis was placed on organizing and optimizing SQL code for quick searches and scalability.

Query Functionality: The application featured a query displaying the top 10 players in the 2021 Masters tournament, highlighting the database's accuracy.

User Interface Design: The application provided features for filtering search results by player, tournament, or date, enhancing user experience.

Outcomes and Learning

The project offered practical experience in database design, SQL scripting, and data management, enhancing skills in database administration and data analysis. It demonstrated the importance of user-friendly interface design and efficient data handling in developing effective database applications.

IST 652 - Scripting for Data Analysis

Analyzing Transaction Data for Fraud Detection: An Approach Using Logistic Regression

This report outlines the analysis of a transaction dataset covering a two-day period, focusing on fraud detection. With only 492 out of 284,315 transactions flagged as frauds (0.172% of all transactions), the dataset presents a highly imbalanced nature. The study demonstrates the use of logistic regression in identifying fraudulent transactions and discusses the challenges and insights gained from handling an imbalanced dataset.

Methodology

The analysis involved several key steps:

Data Preparation: A robust scaler was applied to normalize transaction amounts and times, addressing the variance in these features.

Exploratory Data Analysis: Histograms and correlation matrices were utilized to understand the data distribution and relationships between features, identifying those significantly influencing the occurrence of fraud.

Model Development: A logistic regression model was chosen for its simplicity and interpretability.

Results and Discussion

Model Performance: The logistic regression model achieved an accuracy rate of 99.92% on test data. Crucially, the recall rate was 91.8%, indicating the model's effectiveness in identifying instances of fraud.

Importance of Recall: The project emphasizes the significance of recall in fraud detection, where overlooking fraudulent transactions can lead to substantial financial losses. It also discusses the precision-recall trade-off, particularly relevant in datasets with rare instances of fraud.

Learning Outcomes: The project offered practical experience in handling imbalanced datasets, a frequent challenge in data science. It also provided insights into the application of logistic regression for binary classification problems and underscored the importance of evaluating model performance metrics in a business context.

Conclusion

This analysis demonstrates a comprehensive approach to fraud detection using logistic regression, highlighting the analytical skills necessary to extract meaningful insights from complex datasets. The study serves as a practical example of tackling real-world problems in data science, emphasizing the critical balance between model accuracy and the ability to detect rare events like fraud.

World Happiness Analysis: Investigating Socio-economic and Demographic Influences

The World Happiness Analysis project as part of the IST 652 course, aims to explore the impact of various socio-economic, demographic, and other factors on a country's happiness ranking. This comprehensive study integrates data from the World Happiness Report with additional country statistics to examine the determinants of national happiness levels.

Methodology

The project encompasses several phases:

Setup and Data Merging: Merging the World Happiness Report data with additional country statistics.

Data Preprocessing: Cleaning and preprocessing the data, handling missing values, and ensuring consistent data types.

Exploratory Data Analysis: Providing an initial overview of the dataset.

Factor Analysis: Identifying key socio-economic, demographic, and other factors correlating with happiness rankings.

Regional Comparisons: Comparing happiness scores and influential factors across different world regions.

Predictive Modeling: Building a regression model to predict happiness scores based on identified factors, with model performance evaluated using cross-validation.

Results and Analysis

The regression model explains approximately 83.4% of the variability in the "Ladder Score," a measure of a country's happiness. Key predictors identified include "Logged GDP per capita," "Social Support," and "Freedom to make life choices." The analysis reveals that economic factors like GDP per capita and social support play significant roles in determining a country's happiness rank. Additionally, the freedom to make life choices is a crucial contributor to happiness, highlighting the importance of social and political dimensions in well-being.

Conclusion and Future Recommendations

The project concludes with policy implications, emphasizing the importance of economic development, social welfare programs, and individual liberties in enhancing happiness. Future research directions suggested including addressing potential multicollinearity, exploring additional variables to improve the model, and conducting longitudinal studies.

Implications

The World Happiness Analysis project offers valuable insights into the determinants of national happiness, providing guidance for policymakers and leaders. It also sets a foundation for further research in understanding the complex interplay of various factors affecting a country's happiness.

IST 707 - Applied Machine Learning

Handwriting Recognition through Machine Learning: An Analytical Study

This report aims to recognize handwritten numbers by evaluating and comparing various machine learning methods. The primary focus was to understand the strengths and weaknesses

of each method to determine the most effective approach for this specific task of handwriting recognition.

Methodology

The Kaggle digits dataset, consisting of separate training and testing datasets, was utilized for this analysis, with an emphasis on the training dataset. Cross-validation was employed as the evaluation method. The data, loaded in CSV format, underwent processing where the "label" column was converted into a factor variable.

Model Evaluation and Results

Performance Metrics: The models were assessed using several metrics. The sensitivity for classifying the digit '0' was 78.82%, and the specificity was 97.31%. The overall model accuracy had a 95% confidence interval ranging from 94.28% to 96.14%.

Kappa Statistic: A key metric in the evaluation was the Kappa statistic, with the model achieving a value of 0.9475, indicating a high level of agreement between the model's predictions and the actual values.

Comparative Analysis: Among various models, one model exhibited superior accuracy. However, Naive Bayes and Decision Tree models were noted for their faster training times.

Further Evaluation: The study suggests the need for additional metrics like precision, recall, and the F1 score for a comprehensive understanding of model performance. The importance of feature analysis was also highlighted for improving model performance.

Conclusion

The analysis presents a moderate agreement on the model's ability to perform class-wise recognition of handwritten numbers. While certain models demonstrated high accuracy, the study emphasizes the need for a thorough evaluation to fully understand each model's strengths and limitations. This comprehensive approach helps in identifying the most effective machine learning method for handwriting recognition.

Benjamin J. Heindl - 3595366618
M.S. Applied Data Science

Multimodal Exploration of Mass Killings: A Machine Learning Approach

This study presents a multimodal exploration of mass killings, employing machine learning and text mining techniques to analyze a dataset derived from narrative summaries of such incidents. The focus is on understanding patterns and classifications within the data, using various statistical and machine learning models.

Methodology

The dataset, comprising 553 samples with 167 predictors, was prepared using text analysis methods like Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TFIDF) weighting to transform narrative summaries into a tokenized structure. Data exploration included Association Rule Mining and k-means Cluster Analysis, with the latter indicating a total within-cluster sum of squares of 133.4175.

Experiments and Results

Model Training: The study evaluated several models:

Conditional Inference Tree (CTree): Achieved 59.09% accuracy and a Kappa statistic of 0.38.

Decision Tree (CART): Recorded around 58.39% accuracy and a Kappa of 0.33.

C5.0: Demonstrated 67.21% accuracy and a Kappa of 0.50.

Support Vector Machines (SVM): Yielded 61.09% accuracy and a Kappa of 0.42.

k-Nearest Neighbors (kNN): Reported 56.03% accuracy and a Kappa of 0.37.

Random Forest: Showed the best performance with 67.92% accuracy and a Kappa of 0.51.

Model Tuning: The Random Forest model was further optimized, achieving the best performance with an mtry value of 52, resulting in approximately 69.02% accuracy and a Kappa of 0.53.

Comparative Analysis: A detailed comparison of all models indicated that Random Forest outperformed others in terms of accuracy and Kappa.

Time Analysis: The study also includes an analysis of the time taken for each model's training, highlighting the computational efficiency of the algorithms.

Visualizations and Further Analysis: Various plots and visual representations were used to aid in understanding the models' performances and the data structure.

Conclusion

This comprehensive analysis employs a rigorous approach to examine the patterns and classifications in the context of U.S. mass killings. By leveraging multiple machine learning techniques and text mining, the study provides valuable insights into the data structure and the effectiveness of different models in this domain.

IST 664 - Natural Language Processing

Comparative Analysis of Sentiment Classification in Movie Reviews Using Machine Learning and BERT

This report presents a comprehensive analysis of sentiment classification in movie reviews, employing a range of machine learning techniques and the advanced BERT (Bidirectional Encoder Representations from Transformers) model. The primary objective was to categorize sentiments in movie reviews into positive, negative, or neutral categories.

Methodology

The methodology incorporated Natural Language Processing (NLP) techniques and machine learning algorithms such as Naive Bayes, SVM (Support Vector Machine), Logistic Regression, and the BERT model. Data preprocessing involved sentiment scoring using VADER (Valence Aware Dictionary and sEntiment Reasoner), and Exploratory Data Analysis (EDA) was conducted to examine sentiment distributions and review lengths.

Results

Data Preprocessing and Analysis: The average sentiment score was slightly positive at 0.0427, with most reviews falling in the neutral to slightly negative/positive range.

Model Performance:

Naive Bayes: Achieved accuracies of 0.5133 using Bag of Words (BoW), 0.5254 with Term Frequency-Inverse Document Frequency (TF-IDF), and 0.5197 with VADER-based features.
SVM: Recorded accuracies of 0.5220 (BoW) and 0.5485 (TF-IDF).
Logistic Regression: Demonstrated accuracies of 0.5208 (BoW) and 0.5534 (TF-IDF).
BERT Analysis: Predominantly identified '5 stars' ratings, indicating a large portion of highly positive reviews in the reduced data subset.

Exploratory Data Analysis (EDA): Revealed that neutral sentiments (Sentiment 2) were most common, and reviews with extreme sentiments tended to be longer.

Conclusion

The analysis concluded that TF-IDF features generally enhanced model performance across different algorithms, with Logistic Regression using TF-IDF emerging as the most effective approach. The use of the BERT model highlighted its capability in capturing nuanced sentiments, underscoring its potential in advanced NLP applications. This study illustrates the effectiveness of various machine learning models and NLP techniques in the sentiment analysis of movie reviews, offering valuable insights into sentiment spectrums and the efficacy of different analytical methodologies.

IST 691 - Deep Learning in Practice

Applying Deep Learning for Sarcasm Detection in Text

This report explores the use of Long Short-Term Memory (LSTM) neural networks in detecting sarcasm in textual content. Utilizing the Onion vs HuffPost headlines dataset, which comprises sarcastic and non-sarcastic headlines, the research aims to evaluate the effectiveness of LSTM models in sarcasm detection.

Methodology

The methodology involves the initial setup of necessary libraries and the loading and preprocessing of the dataset for training and testing purposes. The LSTM neural network, comprising an embedding layer and three LSTM layers, is constructed. The model's architecture is detailed, revealing a total of 2,245,029 trainable parameters.

Experiments and Results

Initial LSTM Model: The first LSTM model is trained and tested, achieving an accuracy of 83.42%.

GloVe Embeddings: The study then incorporates GloVe (Global Vectors for Word Representation) embeddings. A new LSTM model using GloVe embeddings in the input layer is constructed, trained, and evaluated, achieving an improved accuracy of 85.61%.

Word Analogies and Biases: Word analogies are explored using GloVe embeddings. The study examines biases in word embeddings by analyzing associations between gender and occupations, revealing traditional gender roles in the embeddings.

Attention Mechanisms: The limitations of LSTMs in processing long sequences are acknowledged. Attention mechanisms are introduced to improve focus on specific parts of input sequences, particularly in sequence-to-sequence translation models, enhancing context maintenance and translation accuracy for longer sentences.

Conclusion

The research demonstrates the potential of LSTM neural networks in sarcasm detection, highlighting the effectiveness of GloVe embeddings in enhancing model accuracy. The exploration of word analogies and biases in embeddings provides insights into underlying representations in NLP models. The study also suggests attention mechanisms as a valuable solution for handling long sequences, thereby extending the applicability of LSTM models in more complex NLP tasks.

Conclusion

This collection of diverse projects, spanning from database management to deep learning, illustrates a profound journey through the multifaceted landscape of data science. Each project, whether it be the meticulous development of a database application in "SQL Tee Mates" or the intricate machine learning models applied in "Handwriting Recognition" and "Multimodal Exploration of Mass Killings," demonstrates a deep understanding of theoretical concepts and their practical application. The ability to draw meaningful insights from complex datasets, as seen in the "World Happiness Analysis" and "Comparative Analysis of Sentiment Classification," reflects an advanced level of analytical acumen. Furthermore, the "Applying Deep Learning for Sarcasm Detection" project highlights the innovative edge of current data science methodologies. These projects collectively represent not just academic achievements but also a readiness to tackle real-world challenges in various domains, showcasing skills in data analysis, machine learning, and natural language processing. They exemplify the capability to transform data into actionable insights and strategic decisions, underpinning the essential role of data science in contemporary problem-solving and decision-making processes.