Emily Gelchie

Project Brainstorming

Project 1 Idea: Sentiment Analysis for McDonald's Review Rating Prediction Using Python

Project Goals: In this data science endeavor, I am setting out to construct a cutting-edge predictive model using Python that will accurately forecast McDonald's restaurant ratings from textual customer reviews. I aim to harness the nuances of language within these reviews, transforming them into quantitative sentiment scores that will serve as reliable predictors for the establishment's star ratings. The successful completion of this project will offer McDonald's critical insights into customer sentiment, enabling them to address service quality and enhance customer satisfaction proactively.

Techniques and Python Libraries: My approach will be deeply rooted in Python's rich data science landscape. I plan to utilize natural language processing (NLP) libraries such as NLTK and spaCy to perform linguistic preprocessing, shaping the unstructured review text into a structured format ripe for analysis. Sentiment quantification will be tackled through sentiment-specific models like VADER or by employing word embeddings with neural networks in TensorFlow/Keras to capture the subtleties of human emotion expressed in text.
For the predictive modeling aspect, I'll leverage scikit-learn's comprehensive suite of algorithms, experimenting with various classifiers like Naive Bayes and Support Vector Machines, not to mention the robust ensemble methods such as RandomForest and Gradient Boosting Machines.

Cross-validation techniques will be employed to discern the most effective model, followed by rigorous hyperparameter tuning to refine its predictive prowess.

To add a geographical dimension to the analysis, I'll use Geopy for reverse geocoding, enriching the dataset with demographic and socioeconomic variables that may correlate with the ratings, such as county-level income averages or population density.

Tentative Timeline with Technical Detail:

- Week 1: Data Preprocessing
    - I'll begin by cleaning and preparing the dataset using Pandas, ensuring data quality and consistency.
    - To identify patterns and anomalies, I will conduct an initial exploratory data analysis (EDA), shown with Matplotlib and Seaborn.
- Week 2: Sentiment Analysis Implementation.
    - I will execute NLP preprocessing and sentiment scoring, extracting features instrumental for the subsequent predictive modeling.
- Week 3: Model Training and Selection
    - This week will see me building, training, and evaluating various classification models, selecting the most promising based on precision and recall, among other metrics.
- Week 4: Integration of Geospatial Data
    - I'll perform reverse geocoding to augment the dataset with geospatial features and incorporate additional data to provide a more comprehensive analysis.
- Week 5: Final Model Training and Documentation

- The final week will involve the completion of model training and the drafting of a detailed report. I will utilize Jupyter Notebooks to present the analysis and results coherently.

Supplementary Datasets and Integration:

I plan to supplement the primary dataset with external data sources, including U.S. Census Bureau datasets, which I will access through APIs or direct downloads. These datasets will be thoughtfully chosen and integrated using Python's data manipulation capabilities to enhance the predictive accuracy of my model.

Through this project, I aim to deliver not merely a predictive model but a sophisticated analytical tool that will enable McDonald's to delve into the relationship between customer feedback and service ratings, fostering an environment of continual improvement and customer-centric decision-making.

Project 2 Idea: Neural Network Analysis of Factors Influencing Song Popularity on Spotify

Project Goals: My project aims to demystify the elements contributing to a song's popularity on Spotify by developing a neural network model. This model will analyze various song attributes to predict their potential popularity, focusing mainly on the impact of artist collaborations. By identifying the characteristics that correlate with higher streaming numbers and listener engagement, I aim to provide insights that could guide artists and producers in creating successful music collaborations.

Techniques and Python Libraries: I will employ Python's robust data science and machine learning libraries. My neural network model will be constructed using TensorFlow and Keras, which offer the flexibility to design simple and complex neural architectures tailored for tabular data like the Spotify dataset. The model will ingest a range of input features, including but not limited to artist collaboration details, genre classifications, and audio features such as tempo, key, and danceability.

Feature engineering will be a critical component of the preparatory work, where I'll transform categorical data using one-hot encoding and scale numerical features for optimal neural network performance. I plan to utilize Pandas for data manipulation, sklearn for preprocessing, and Matplotlib and Seaborn for visualizing the data and model performance metrics.

Tentative Timeline with Technical Detail:

- Week 1: Data Exploration and Preprocessing
    - I'll conduct thorough EDA with Python tools to understand the dataset's structure and distributions.
    - Data preprocessing will involve handling missing values, encoding categorical variables, and normalizing numerical features.
- Week 2: Feature Engineering and Initial Model Building
    - This week will be dedicated to crafting a feature set that accurately represents the factors influencing song popularity.
    - I will build a preliminary neural network model and establish a baseline for performance.
- Week 3: Model Optimization and Training

        - To prevent overfitting, I'll experiment with different neural network architectures, including deep learning techniques and regularization methods.
        - The model will be trained on a subset of the data, using techniques like k-fold cross-validation to ensure its generalizability.
- Week 4: Model Refinement and Evaluation
    - Based on the previous week's results, I will refine the model, tuning hyperparameters and potentially incorporating additional layers or neurons.
    - Evaluation will be based on accuracy, precision, recall, and the area under the ROC curve.
- Week 5: Final Evaluation and Documentation
    - I will train the final model on the complete dataset.
    - I will compile a comprehensive report detailing the methodology, model architecture, performance, and insights derived from the analysis.

Supplementary Data and Integration: Besides the provided Spotify dataset, I may consider integrating external datasets that offer listener demographics or genre popularity trends over time. Such datasets provide a richer context for the model, potentially improving its predictive accuracy.

By the end of this project, I aspire to have developed a neural network model that not only predicts song popularity but also uncovers the complex interplay of factors that contribute to a track's success on Spotify. This could be a valuable tool for artists, record labels, and music marketers to strategically plan their releases and collaborations.