Final Project

K.N.Toosi University of Technology
Introduction to Data Mining

Fall 2024

Part I
Sales Prediction

Dataset

The dataset is available for download on the course website.

Dataset Description

- **Store_id**: A unique identifier for each store.
- **RetailType**: Category or type of retail store (e.g., grocery, clothing, electronics).
- **Stock variety**: Range of products offered (e.g., basic, extended, premium).
- **DistanceToRivalStore**: Distance from this store to its nearest rival.
- **RivalOpeningMonth**: Month when a rival store opened nearby.
- **RivalEntryYear**: Year when a nearby rival store entered the market.
- **ContinuousBogo**: Whether a "Buy One Get One" (BOGO) offer is active.
- **ContinuousBogoSinceWeek**: Number of weeks since BOGO started.
- **ContinuousBogoSinceYear**: Year when BOGO started.
- **ContinuousBogoMonths**: Total months BOGO has been active.
- **DayOfWeek**: The day of the week sales data was recorded.
- **Date**: Specific date of sales data.
- **Sales**: Store's total sales made on the given day.
- **NumberOfCustomers**: Number of customers visiting the store.
- **Is_Open**: Whether the store was open that day (1 for open, 0 for closed).
- **BOGO**: Whether a BOGO offer was active (1 for active, 0 for not active).
- **Holiday**: Whether the day was a recognized holiday (1 for holiday, 0 for non-holiday).

Task

1. **Load Dataset:**
    - Read training data from CSV, selecting relevant columns. Consider that some columns may not be useful for your analysis and can be omitted.
2. **Load and Merge Store Data:**
    - Read the stores data from another CSV file to get additional information about each store.
    - Combine the training and store data based on the store id column to create a comprehensive dataset.
3. **Train & Test Data:**
    - Divide 70% of the training examples into the training set and use the remaining 30% as the test set. Select the first 70% of the examples in chronological order, as we aim to evaluate our models on their ability to extrapolate to dates beyond the training range.
4. **Preprocess Data:**
    - Replace the missing values in the 'DistanceToRivalStore' column with the median of the existing values, and set the remaining missing values to zero. Feel free to modify this for better approaches if you prefer.
    - Extract 'Year', 'Month', 'Day', and 'WeekOfYear' from the 'Date' column, then remove the 'Date' column. Utilize the `pd.to_datetime` function and its attributes for easy handling.
    - Remove the 'Customers' column since it is not available during testing.
    - Standardize features using `StandardScaler`
5. **Prepare Data for Modeling:**
    - Separate the features (X) and the target variable (y), which is total_sales.
6. **Train and Evaluate Models:**
    - Train Linear Regression and Random Forest Regressor; Evaluate their performances on the test data.
7. **Feature Selection & Importance:**
    - Utilize `feature_importances_` from the Random Forest model. Which features were identified as the most important by this model?

Part II
Sentiment Analysis

Dataset

The dataset is available for download on the course website.

Task

1. **Load and Preprocess Data:**
    - Load dataset and preprocess reviews (e.g., lowercasing, removing stopwords, punctuation).
2. **Vectorize Reviews:**
    - TF-IDF: `from sklearn.feature_extraction.text import TfidfVectorizer`
    - Word2Vec: `from gensim.models import Word2Vec`
    - BERT: `from sentence_transformers import SentenceTransformer`
3. **Hyperparameter Tuning for Classification Models:** [1]
    - Conduct a grid search on each model to identify the optimal hyperparameters, utilizing the F1 score for evaluation.
        - Logistic Regression
        - Random Forests
        - K-Nearest Neighbors

Note

Any attempt to use AI tools for generating the code is strictly prohibited. Students will be asked to present and explain their code during a class session.

[1] Due to the potential data imbalance caused by splitting reviews at a threshold of 3, we compare three different algorithms to ensure robustness.