ITM 760 - Big Data Analytics

Final Group Assignment

Group Members:

Biratu, Michael	501052498
Gupta, Rohan	501044578
In, Young Jae	500902059
Torres, Diego	501143240

Toronto Metropolitan University

Introduction

- To gain hands on experience using Python to build models from real-world datasets.
- To evaluate different data mining algorithms in terms of accuracy & run-time
- Two datasets
  - Online News Popularity
  - Online Shoppers Purchasing Intention

Objectives

- **Online News Popularity Dataset:** Predict the number of article shares (popularity) as a regression task using features like title length, content length, and sentiment indicators; evaluated with MSE, RMSE, and MAE.
- **Online Shoppers Purchasing Intention Dataset:** Classify whether a user session generated revenue, using attributes like bounce rates, browser, and region; evaluated with accuracy, precision, and recall.
- **Algorithms Applied:** Use Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), and Decision Tree for both regression and classification tasks.

Toronto Metropolitan University

Dataset #1 - Online News Popularity

Toronto Metropolitan University

Dataset 1: Online News Popularity

Exploratory Data Analysis & Data Cleaning

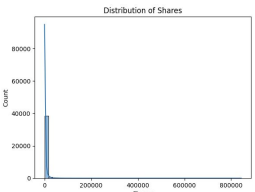


Figure 1. Distribution of Shares

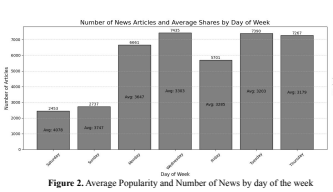


Figure 2. Average Popularity and Number of News by day of the week

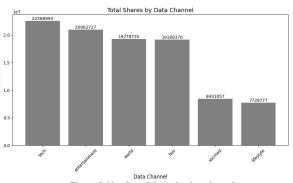


Figure 3. Number of shares by data channel

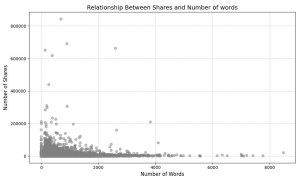


Figure 4. Number of shares versus number of words in an article

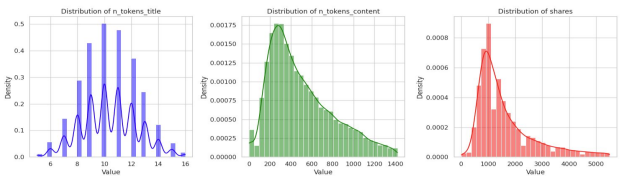


Figure 6. Updated distributions for n_tokens_title, n_token_content, shares

1. Examine data distribution to identify outliers
2. Explore relationship between “popularity of news articles” vs “day of publication”
3. Explore correlation between popularity vs data channel (tech, entertainment, social media, etc)
4. Assess if length of article impacts its popularity
5. Compare popularity of negative and positive polarity

Toronto Metropolitan University

Data Cleaning

Summary of Data Exploratory

- **Target Features:** The distribution of shares shows significant outliers, with a few articles achieving extreme popularity; most remain within a similar range.
- **Key Influencing Factors:** Weekend articles are more popular despite fewer publications, shorter articles receive more shares, and negative polarity scores correlate with higher popularity.
- **Data Channel Trends:** Social Media and Lifestyle channels are less popular compared to Tech, Entertainment, World, and Business channels.

Cleaning Actions

1. **Cleaning:** Removed null values, duplicates, and outliers using pandas and the Interquartile Range (IQR) method for key attributes like "*n_tokens_title*", "*n_tokens_content*", and "*shares*".
2. **Dataset Reduction:** After cleaning, the dataset size was reduced to 28.5k records, ensuring a refined and consistent dataset.
3. **Data Splitting:** Utilized scikit-learn to randomly split the cleaned data into 70% training and 30% testing subsets for reliable model evaluation.

Toronto Metropolitan University

Machine Learning Algorithms Results

Evaluated SVM, KNN, and Decision Tree regression models using all 60+ features and a refined set of 19 key features to balance accuracy, overfitting prevention, and computational efficiency, with performance assessed via MSE, RMSE, and MAE.

1. **Support Vector Machine (SVM):** Achieved the **lowest RMSE** (**1,074.25**) but had the longest training time (33.97 seconds with all features); potential overfitting due to dataset non-linearity.
2. **K-Nearest Neighbors (KNN):** Fastest training time (0.021 seconds with all features), consistent RMSE (1,111.46), but limited by high-dimensional data.
3. **Decision Tree:** Struggled with overfitting; best performance with selected features (RMSE: 1,471.50), but overall poorest among the models evaluated.

Model Comparison

Two versions of each algorithm were trained—one using all features and another with 19 selected features, with training times doubling for full-feature models.

- **Performance Rankings:** SVM achieved the best overall results, followed by KNN, while the Decision Tree performed the worst across all metrics (MSE, RMSE, MAE).

Toronto Metropolitan University

Dataset #2 - Online Shoppers Purchasing Intention

Toronto Metropolitan University

Dataset 1: Online News Popularity

Exploratory Data Analysis & Data Cleaning

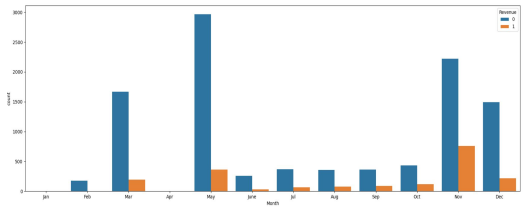


Figure 1. Revenue per Month

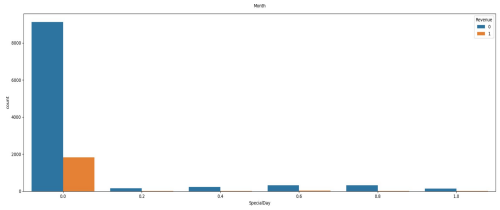


Figure 2. Relationship between Holiday & Distance of Time

- Examine data distribution to identify outliers
- Explore relationship between “month” vs “revenue”
- Explore correlation between distance of time between the date a person purchases vs data channel

- **Data Cleaning:** Removed duplicates, converted "Revenue" and "Weekend" columns to binary integers, and transformed "Month" and "VisitorType" into separate binary columns for clearer analysis.
- **Refinement:** After cleaning and restructuring, the dataset size was reduced to 12,329 records, with all attributes converted to binary integers for improved machine learning performance.
- **Insights:** Separated months enabled better visualization and understanding of revenue trends throughout the year, highlighting the impact of time on purchasing behavior.

Toronto Metropolitan University

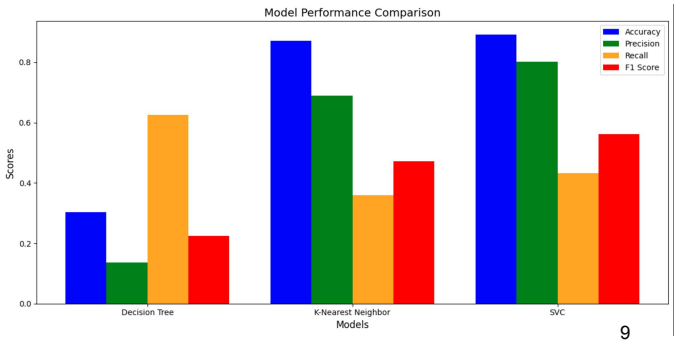
Machine Learning Algorithms Results

Used SVM, KNN, and Decision Tree to classify whether revenue was generated, evaluating performance with accuracy, precision, and recall.

1. **Support Vector Machine (SVM):** Achieved high accuracy (89.16%) and precision (80.13%), but low recall (43.20%) indicates missed positives; F1 score: 0.561.
2. **K-Nearest Neighbors (KNN):** Good accuracy (87.11%) and precision (68.95%), but recall (35.88%) is low, resulting in a below-average F1 score of 0.472.
3. **Decision Tree:** Poor performance with low accuracy (30.34%) and precision (13.63%); high recall (62.59%) due to over-predicting positives; F1 score: 0.224.

Model Comparison

SVM outperformed KNN in precision, recall, and F1 score, while the Decision Tree significantly underperformed; both SVM and KNN achieved similar accuracy (89% and 87%, respectively).



Toronto Metropolitan University

Recommendations

Dataset #1

Use the KNN model with selected features for its efficiency, minimal risk of overfitting, and competitive performance; SVM with selected features is a secondary choice for larger datasets despite longer training times.

Dataset #2

After evaluating SVM, Decision Tree, and KNN, SVM is the best model for the dataset, achieving the highest accuracy (89.16%), strong precision (80.13%), and a balanced performance between precision and recall, making it the most reliable classifier overall.

Toronto Metropolitan University