Postgraduate Diploma in Science in Data Analytics 2020

Data Mining & Machine Learning 1

Project Proposal

Aloysious Brhanavan (x19206984)

Lecturer: Dr. Pierpaolo Dondio

Contents

1 Motivation

This project explores 3 different datasets, two classification and one regression problem using five machine learning techniques.

Businesses rely heavily on data to make informed decisions. However, bulk of such useful data comes in the form of unstructured text data from emails, social media, customer feedback, surveys and articles. Manually shifting through mountains of text data to gain insights is both tedious and time consuming. Sentiment analysis, a branch of Natural Language Processing (NLP) helps to identify sentiment given a text. The first dataset *Rotten Tomatoes reviews* involves classifying the audience opinion as positive or negative from the reviews.

Businesses solicit feedback through customer surveys in order to evaluate their service against customer expectations. This helps assess how happy are customers with various aspects of products and service and measure customer satisfaction. The second dataset, Airline Passenger satisfaction involves predicting overall customer satisfaction of a particular journey using factors such as age, gender, type of travel and customer feedback regarding different aspects of the airline's services. One of the objectives of this analysis is to identify the factors that matter the most to airline passengers so airlines could focus on these to achieve a better overall passenger satisfaction.

For those who are taking the plunge and getting married one of the biggest expenses is buying a diamond for the engagement ring. Bluenile is one of the largest online specialist engagement ring retailers in the US and operates as a marketplace. Using the dataset *Diamond prices* sourced from Bluenile, I will employ machine learning to predict price of a diamond based on its characteristics. I hoped to demonstrate bargains can still to be found.

2 Research Questions

Dataset 1 – Rotten Tomato views: How accurately can we predict whether the sentiment of a movie review is positive or negative?

Dataset 2- Airline passenger satisfaction: How well can we predict whether an airline passenger is satisfied or not satisfied on a particular journey using variables such as age, gender, travel type and passenger feedback on different aspects of the service? What factors have the biggest impact on airline passenger satisfaction?

Dataset 3 – Diamond prices: How accurately can we predict price of a diamond based on its characteristics (such as carat, colour, clarity, cut, depth etc)?

3 Initial Review

1. The 2011 research paper “Learning Word Vectors for Sentiment Analysis” [1] from Stanford university introduced the now popular IMDB 50K movie dataset. The authors demonstrated Linear SVM classifier showed superior performance compared to other methods and achieved a highest accuracy of 88.33% in classification when combined with Bag-of-words representation. However, this paper only considered highly polarized reviews and excluded neutral reviews of rating between 5 and 7 stars out of 10 from the dataset.

2. “SACPC: A framework based on probabilistic linguistic terms for short text sentiment analysis” [2] paper published in April 2020 compared the performances of Naive Bayes, SVM and Logistic Regression models on a balanced single sentence movie review dataset of size 10622 collected from Rotten Tomatoes website. Naive Bayes, SVM and Logistic regression with bag-of-words representation achieved precision of 0.75, 0.75 and 0.76 and recall of 0.78, 0.76 and 0.77 respectively on Rotten Tomatoes movie review dataset.

3. “Sentiment Analysis of Twitter Data: A Survey of Techniques” [3] compared the performances of various machine learning models including the effect of using ngrams. The paper concluded that machine learning models such as SVM and Naive Bayes tend to produce the highest accuracy for text sentiment classification.

4. The 2019 paper “Comparison of the efficiency of Machine Learning algorithms on Twitter Sentiment Analysis of Pathao” [4] performed a comparative efficiency analysis of 3 machine learning algorithms on classifying Twitter sentiment. SVM achieved the highest accuracy of 82.3% followed by Naive Bayes classifier with accuracy scores of 79.3%.

5. The 2016 research paper “Comparison of 14 different families of classification algorithms on 115 binary datasets” [5] compared the performance of various classification algorithms on 115 real life binary datasets and concluded the 3 best classifiers were Random Forest, SVM with Gaussian kernel and Gradient Boosting.

6. “An empirical comparison of ensemble methods based on classification trees” [6] paper published in 2005 compared the accuracy of several ensemble methods using 14 data sets. It found the best overall results were obtained with Random Forest. Moreover, Random Forest proved the most robust against noise too.

7. “Machine Learning Algorithms for Diamond Price Prediction” [7] published in March 2020 compared the predictive power of several machine learning algorithms

including liner regression, Random forest regression, polynomial regression, Gradient descent and Neural network. It ranked Random Forest as the best estimator for predicting diamond prices.

8. The 2019 paper “Gold and Diamond Price Prediction Using Enhanced Ensemble Learning” [8] used a Kaggle diamond prices dataset of 53940 samples and 10 features such as carat, colour, clarity cut etc to predict diamond prices. Random Forest and Linear regression achieved R-squared values of 0.973 and 0.874 respectively.

4 Data Sources

4.1 Dataset 1: Rotten Tomatoes Movie Reviews (RTMR)

This dataset was sourced from Rotten Tomatoes. First the links to 142 summer movies released between 2017 and 2019 were programmatically extracted from three Rotten Tomatoes editorial article web pages using the python package Beautiful Soup. Then using these URLs, the corresponding movie ids were downloaded from Rotten Tomatoes. Finally, the Rotten Tomatoes API was queried using the movie ids to obtain the audience review and the corresponding star rating for each movie. Because some movies, especially blockbusters tend to have substantially more reviews than others, the maximum reviews per movie was limited to 2500.

Data source: https://www.rottentomatoes.com/

Dimensions: 169,669 rows and 2 columns

Variables:

- Review: Text
- Star Rating: between 0.5 stars and 5 stars

Predicting: Sentiment (derived from star rating, rating of 2.5 stars and below is negative and rating of 3 stars and above is positive).

4.2 Dataset 2: Airline Passenger Satisfaction (APS)

Data source: https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction

Dimensions: 129,880 rows and 25 columns, split into training set of size 103,904 rows and test set of size 25,976 rows

Variables:

- Id: unique identifier for each journey
- Nominal: Gender, Customer type, type of travel, Class
- Ordinal: The following variables are based on customer feedback on various aspects of the journey, scored 0 to 5.
  Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Check-in service, Inflight service, Inflight service
- Numerical Continuous: Age, Flight Distance, Departure Delay in Minutes, Arrival Delay in Minutes

Predicting: Customer Satisfaction (satisfied or not satisfied). 56,428 satisfied and 73,452 not satisfied.

4.3 Dataset 3: Diamond prices (DP)

This dataset was sourced from online diamond retailer Bluenile.com using their public API.

Data source: http://bluenile.com/

Dimensions: 21,245 rows and 14 columns

Variables:

- Id : unique diamond ID
- Nominal: ShapeName
- Ordinal: colour, clarity, cut, culet, fluorescence, polish, symmetry
- Numerical discrete: lxwRatio
- Numerical continuous: carat, depth, table

Predicting: Price

5 Machine Learning Methods

5.1 Naive Bayes

Naive Bayes classifier is a machine learning algorithm that is based on Bayes Theorem. It makes the naive assumption that effect of a particular feature is independent of other features.

Dataset 1 – RTMR uses Multinomial Naive Bayes for text classification.

5.2 Linear SVM

SVM uses kernel trick technique to transform input data from lower dimensional space into a higher dimensional space in order to find maximum marginal hyperplanes that best divide the dataset into classes. Linear SVM uses a linear kernel.

Dataset 1 -RTMR uses SVM for text classification.

5.3 Random Forest

Random Forest is an ensemble leaning method that uses decision trees are base estimator. Each estimator is trained on an independent bootstrap sample and each tree considers only a random subset of original features. This increases diversity in the forest which results in more robust predictions. In classification each tree votes and the final prediction the majority decision. In regression the average of all tress is the final prediction.

Dataset 1 -RTMR and Dataset 2- APS use Random Forest classifiers

Dataset 3- DP uses Random Forest Regressor to predict diamond prices

5.4 Gradient Boosting

Gradient Boost is an ensemble learning method that combines many weak learners to form a strong learner. Gradient boosting is an iterative method where each estimator in the ensemble is trained using its predecessors' residual errors.

Dataset 2 – APS uses Gradient Boosting Classifier for classification.

Dataset 3- DP uses Gradient Boosting Regressor to predict diamond prices.

5.5 Ada Boost

Ada Boost or Adaptive boost is an ensemble learning method that combines many weak learners to form a strong learner. Adaboost is an iterative method where each subsequent estimators give more attention to the errors of its predecessor by adjusting the weights of the training instances.

Dataset 2- APS uses AdaBoost classifier.

Dataset 3- DP uses AdaBoost regressor to predict diamond prices

6 Evaluation Methods

6.1 Accuracy

The accuracy of a classifier is the percentage of instances that is correctly labelled by the classifier.
Dataset 1 – RTMR and Dataset 2- APS use this metric for evaluation.

6.2 Precision

Precision measures what proportion of positive predictions are actually correct.
Dataset 1 – RTMR and Dataset 2- APS use this metric for evaluation.

6.3 Recall

Recall measures what proportion of actual positives are correctly labelled.
Dataset 1 – RTMR and Dataset 2- APS use this metric for evaluation.

6.4 F1 Score

The harmonic mean of precision and recall.
Dataset 1 – RTMR and Dataset 2- APS use this metric for evaluation

6.5 Kappa

The kappa statistic adjusts the notion of accuracy by also accounting for the possibility that a correction prediction is chance
Dataset 1 – RTMR and Dataset 2- APS use this metric for evaluation

6.6 Confusion Matrix

A confusion matrix summarizes the performance of a classification algorithm as a table.
Dataset 1 – RTMR and Dataset 2- APS use confusion matrix for evaluation

6.7 ROC AUC Curve

The area under the Receiver Operating Characteristics curve at various threshold values for classification problems.
Dataset 1 – RTMR and Dataset 2- APS use ROC AUC curve for evaluation

6.8 Balanced Accuracy

Balanced Accuracy is the average of recall obtained in each class.
Dataset 1 – RTMR uses this metric for evaluation.

6.9 R-squared

R-squared is a goodness of fit measure of a regression model. It represents the proportion of the variance of the target variable that is explained by the independent predictors of the model.
Dataset 3 – DP uses R-squared for evaluation.

6.10 MSE & RMSE

The mean squared error (MSE) measures the average squared difference between the estimated values and actual values.
RMSE is the square root of MSE
Dataset 3 – DP uses MSE and RMSE for evaluation.

6.11 MAE

The mean absolute error (MAE) measures the average absolute difference between the estimated values and actual values.
Dataset 3- DP uses MAE for evaluation.

7 Summary Table

	Dataset1: RTMR	**Dataset2: APS**	**Dataset3: DP**
Data source	Rotten Tomatoes	Kaggle	Bluenile
Rows	169,669	25,976	21,245
Columns	2	25	14
Type of problem	Classification	Classification	Regression
Predicting	Review sentiment (Positive or negative)	Passenger Satisfaction (satisfied or not satisfied)	Diamond Price
ML Methods	Random Forest Naive Bayes Linear SVM	Random Forest Gradient Boost AdaBoost	Random Forest Gradient Boost AdaBoost
Evaluation Methods	Accuracy Balanced Accuracy Precision Recall F1 Score Kappa Confusion Matrix ROC AUC curve	Accuracy Precision Recall F1 Score Kappa Confusion Matrix ROC AUC curve	R-squared MSE RMSE MAE

8 Bibliography

[1] Maas, Andrew & Daly, Raymond & Pham, Peter & Huang, Dan & Ng, Andrew & Potts, Christopher. (2011). Learning Word Vectors for Sentiment Analysis. 142-150.

[2] Song, C., Wang, X., Cheng, P., Wang, J., & Li, L. (2020). SACPC: A framework based on probabilistic linguistic terms for short text sentiment analysis. Knowledge-Based Systems, 194, 105572. doi:10.1016/j.knosys.2020.105572

[3] Kharde, Vishal & Sonawane, Sheetal. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. International Journal of Computer Applications. 139. 5-15. 10.5120/ijca2016908625.

[4] Sajib, Mahamudul & Shargo, Shoeib & Hossain, Md. (2019). Comparison of the efficiency of Machine Learning algorithms on Twitter Sentiment Analysis of Pathao. 1-6. 10.1109/ICCIT48885.2019.9038208.

[5] Wainer, Jacques. (2016). Comparison of 14 different families of classification algorithms on 115 binary datasets.

[6] Hamza, Mounir & Larocque, Denis. (2005). An empirical comparison of ensemble methods based on classification trees. Journal of Statistical Computation and Simulation - J STAT COMPUT SIM. 75. 629-643. 10.1080/00949650410001729472.

[7] Alsuraihi, Waad & Al-hazmi, Ekram & Bawazeer, Kholoud & Alghamdi, Hanan. (2020). Machine Learning Algorithms for Diamond Price Prediction. 150-154. 10.1145/3388818.3393715.

[8] Pandey, Avinash & Misra, Shubhangi & Saxena, Mridul. (2019). Gold and Diamond Price Prediction Using Enhanced Ensemble Learning. 1-4. 10.1109/IC3.2019.8844910.