Text Analysis of Automobile Consumer Reviews

Augustine Joseph

Final Project Report

E-Mail : aujoseph@umail.iu.edu

INFO-I590-34717 – Fall Semester 2015 – Prof. Geoffrey Fox

Indiana University

Introduction

The edmunds.com website provides an excellent platform for car buyers to express their opinions, indicate dissatisfaction with the performance, influence future improvements and rate vehicles on several parameters like performance, comfort, fuel economy, interior design, exterior design, reliability and build quality. These consumer reviews contain text and data, and when analyzed, can generate significant knowledge about vehicles. This paper reports an exploratory data analysis performed on consumer reviews data, to find out the most popular features and most disliked features of a car. Also this paper reports a supervised learning performed on the reviews data to predict the customer sentiment about how favorite is certain vehicle to the reviewer.

**Problem Definitions**

**Background**

Edmunds.com is a web site driven to make automobile buying by providing information, specifications and reviews about all makes and models of automobiles that are sold in the US. This website has become one of the most trusted sites to learn about automobiles before making purchases. It is estimated that about 18 million visitors use it's shopping tools every month to connect with over 9,500 dealer franchises across the U.S. Edmunds.com was named by Maritz Research as one of the most trusted online consumer review sites (Pr, 2014). Many customers who actively engage with this website, have participated in creating a repository of honest consumer reviews. These reviews are mostly sought through channels that can verify the automobile purchase before accepting reviews. The Edmunds.com website now offers application programming interface(API) to retrieve variety of data on automobiles. This API encompasses a wide variety of vehicle-centric datasets including vehicle specifications, consumer ratings and reviews, stock photos, vehicle pricing and maintenance schedules just to name a few. The API has the biggest dataset Edmunds.com has to offer(Edmunds.com, 2015). The Vehicle API can be used to retrieve consumer reviews on automobiles. API calls to the website's developer interface return JSON objects with a wealth of information. An element of this JSON object called "Suggested Improvements" captures the consumer complaints on features of the vehicle that are sub-standard or needing improvements. An exploratory analysis of this element will help us to identify the mostly disliked features of a vehicle. The JSON array element called "Favorite Features" store the customers response to the features that they are happy about to have in their vehicle. An exploratory analysis of this element will help us to identify the most popular features of a vehicle. Another JSON element called "text" stores the overall review of the vehicle. The JSON element "average rating" assigns a number between one to five indicating the customer's level of satisfaction with the vehicle. By analyzing the customer's overall review we could predict customer's level of satisfaction with the vehicle.

**Goals and Objectives**

The goals of this study are to perform: 1) Exploratory data analysis to identify the most popular features and most disliked features of an automobile, 2) A supervised customer sentiment analysis to predict the overall customer satisfaction by analyzing the text data of customer reviews. I have chosen 2008 Honda Accord, a popular sedan for this study since it has received a large number to customer reviews.

The objectives of this study are:

1) Utilization of text mining techniques for an unsupervised learning to discover text patterns in the suggested improvements section of the automobile review forum. With this study I aim to find out six topics of dissatisfaction, the customers have with the vehicle.

2)Application of the same techniques to identify six names of popular features of this vehicle.

3)Supervised learning of the text reviews using multinomial Naive Bayes to predict overall customer satisfaction level with the vehicle.

**Tasks**

**Unsupervised Learning:** This study aims at extracting the major topics that the customers discuss in the suggested improvements section and favorite features section of the automobile consumer reviews. The following tasks were identified to achieve the goals and objectives of this study.

1) Automate the download of multiple customer reviews through API calls to Vehicle API.

2) Parse the JSON objects returned by the API calls to extract relevant elements(text/information) that require analysis.

3) Preprocessing the data to generate meaningful elements of word using tokenizer and lemmatizer.

4) Using term frequency–inverse document frequency(tf-idf) , detect important list of words in a collection or corpus

5) Fit the bag of words in step 4 to a Non-Negative matrix Factorization to generate most popular features and most disliked features of a car.

**Supervised Learning:** For this study if a reviewer has rated, four or more on average ratings of the car, it is considered as a favorite car(favorite=1). If not, the car is considered as not a favorite(favorite=0). With an appropriately trained, calibrated and cross validated model, the study intents to accurately predict whether the car is a favorite car for the reviewer, based on the text analysis of the review. The following tasks were identified to achieve the goals and objectives of this study.

1) Automate the download of multiple customer reviews through API calls to Vehicle API.

2) Parse the JSON objects returned by the API calls to extract relevant elements(text/information) that require analysis.

3) Preprocessing the data to generate meaningful elements of word using tokenizer and lemmatizer.

4) Using term frequency–inverse document frequency(tf-idf) , detect important list of words in a collection or corpus.

5) Determine a set of reviews that will be a member of either a favorite or a not so favorite car.

6) Create a "training corpus" of reviews that have known value of "favorite" associated with them.

7) Train the Naive Bayes classifiers on the training corpus using Python's scikit learn library.

8) Use the above model to label new reviews.

9) Calibrate the model to show the confidence levels of prediction.

10) Validate the classifier against unseen data using cross validation.

**Software Solution**

**List of tools**

A number of software tools and libraries were used to implement the tasks listed above. Python programming has been done using Enthought Canopy version 1.6.1. The following Python libraries were used to implement scientific and machine learning computations: IPython3.2.0, Matplotlib1.4.3, sickit learn 0.17, nltk3.1, numpy1.9.2, panda0.16.2 and scipy0.16. This software has been built on a desktop computer running MS-Windows 7 Professional.

**List of techniques**

**Unsupervised Learning:** To automate the download of multiple customer reviews a URL object has been created that points to edmunds.com API with an embedded apikey. This URL object has been used to generate JSON objects. These JSON objects were used to generate a list of text data for suggested improvements and favorite features. These text data were passed on to a function that can extract six major topics of discussions for suggested improvements and favorite features. This function implements tokenizers and lemmatizers to generate meaningful elements of words. English language stop words were used to filter out common words. Ngrams(1-2) were used to generate meaningful combination of words. After this, the preprocessed text data has been used to generate tf-idf vectors using scikit learn libraries. The term frequency matrix of a corpus of data each for suggested improvements and favorite features has been used with a Non-Negative Matrix Factorization model(Grisel, 2010). This model helps to extract an additive model of the topic structure of the corpus. The output is a list of six topics described with six words or meaningful combinations of words each for suggested improvements and favorite features. Since this implementation has polynomial computational time, 300 sample reviews were used to execute this in a timely manner.

**Supervised Learning:** Similar to the steps followed during unsupervised learning, a URL object has been created with embedded API keys, to point to the edmunds.com vehicle API. JSON objects were created using this URL object. These JSON objects were used to extract a list of text objects for the overall reviews. After preprocessing and filtering stop words, a tf-idf vectorizer from scikit learn library has been used to create a bag of meaningful words and combination of words. This became the independent variable(X) for the Naive Bayes model. The JSON object was also used to extract the average ratings data. If the average rating is below four, the car is considered as, not so favorite and the dependent variable "favorite"(Y) will assume a value of zero. Otherwise the dependent variable "favorite"(Y) will have a value of one. The entire data has been split into training set and testing set using scikit learn library functions(Notebook, 2010). Multinomial Naive Bayes classifiers were used on the training corpus and were used to label the test dataset. The model was calibrated using sigmoid calibration algorithm. The accuracy results were generated.

**Results**

**Unsupervised Learning**

Consumer reviews on 2008 Honda Accord has been used for the unsupervised learning. The results are show in Figure-1 and Figure-2. The most popular features of the 2008 Honda Accord are: interior design, exterior design, heated seat, xm radio, interior room, body style, ride etc. The most disliked features are memory seat, road noise, bluetooth, brake, sound insulation etc. The term gas mileage appears in both popular feature and most disliked features. This means that the model needs additional information to figure out whether gas mileage is a popular feature or a disliked feature. This is a limitation of the current model and may be improved by increasing the ngram size and or by using an entirely different machine learning model.

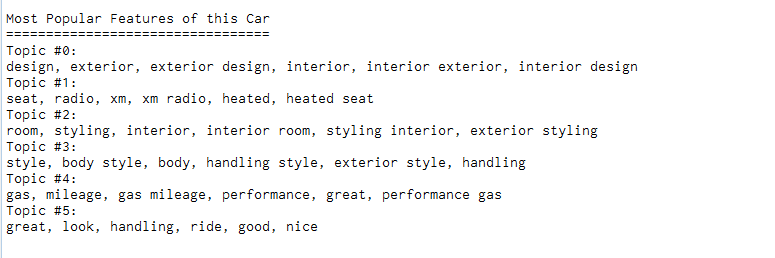


Figure-1: Most popular features of 2008 Honda Accord

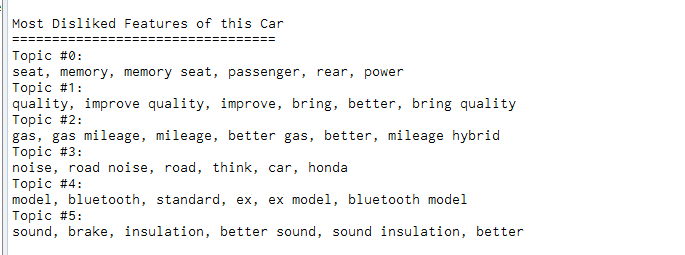


Figure-2 : Most disliked features of 2008 Honda Accord.

**Supervised Learning**

Reviews on 2008 Honda Accord has been used for supervised learning to predict if the car is considered as a favorite vehicle by the reviewer. The dependent variable "favorite"(Y) has a value of one, when the review is positive and it is zero when the review is negative. Multinomial Naive Bayes model has been built to predict the "favorite"(Y) variable. The model has been calibrated using sigmoid calibrator from scikit learn library. The results are illustrated in Figure-3, Figure-4 and Figure-5.

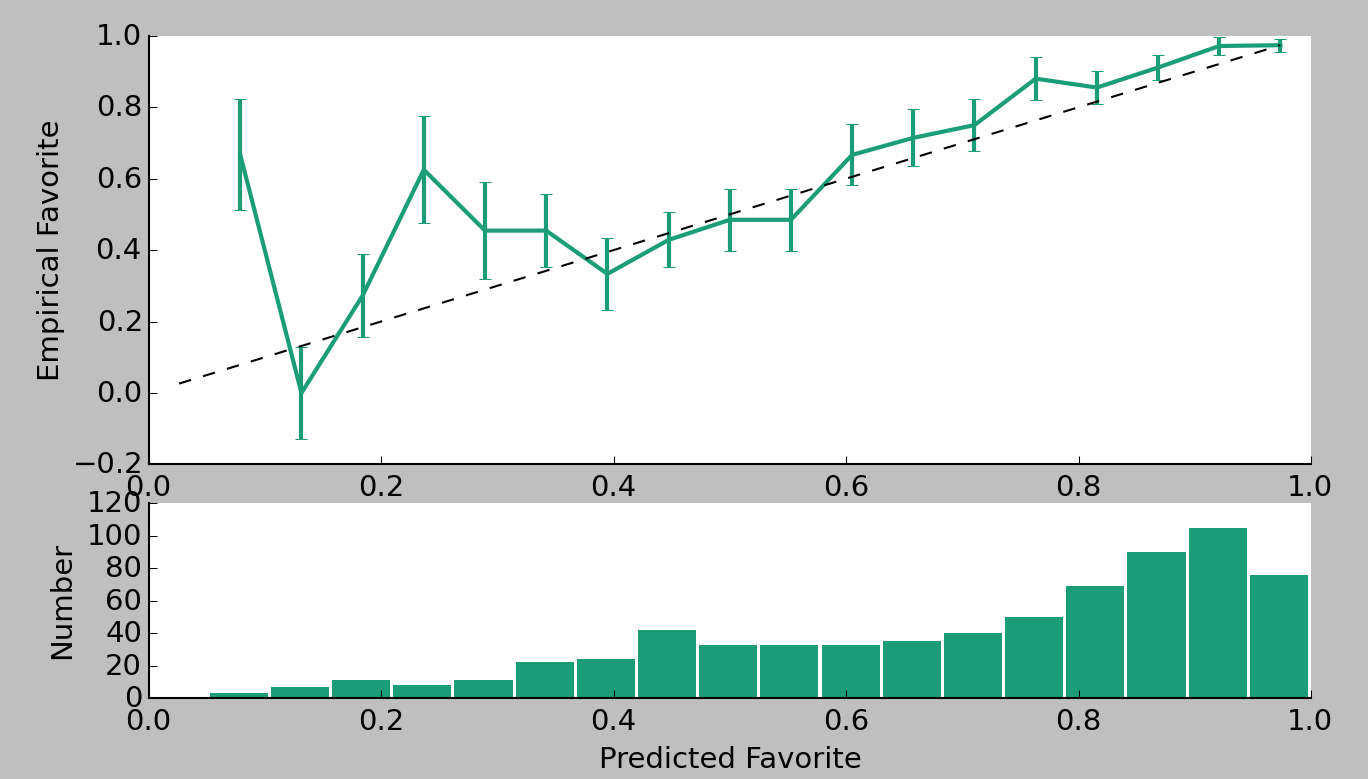


Figure-3 : Calibration Results

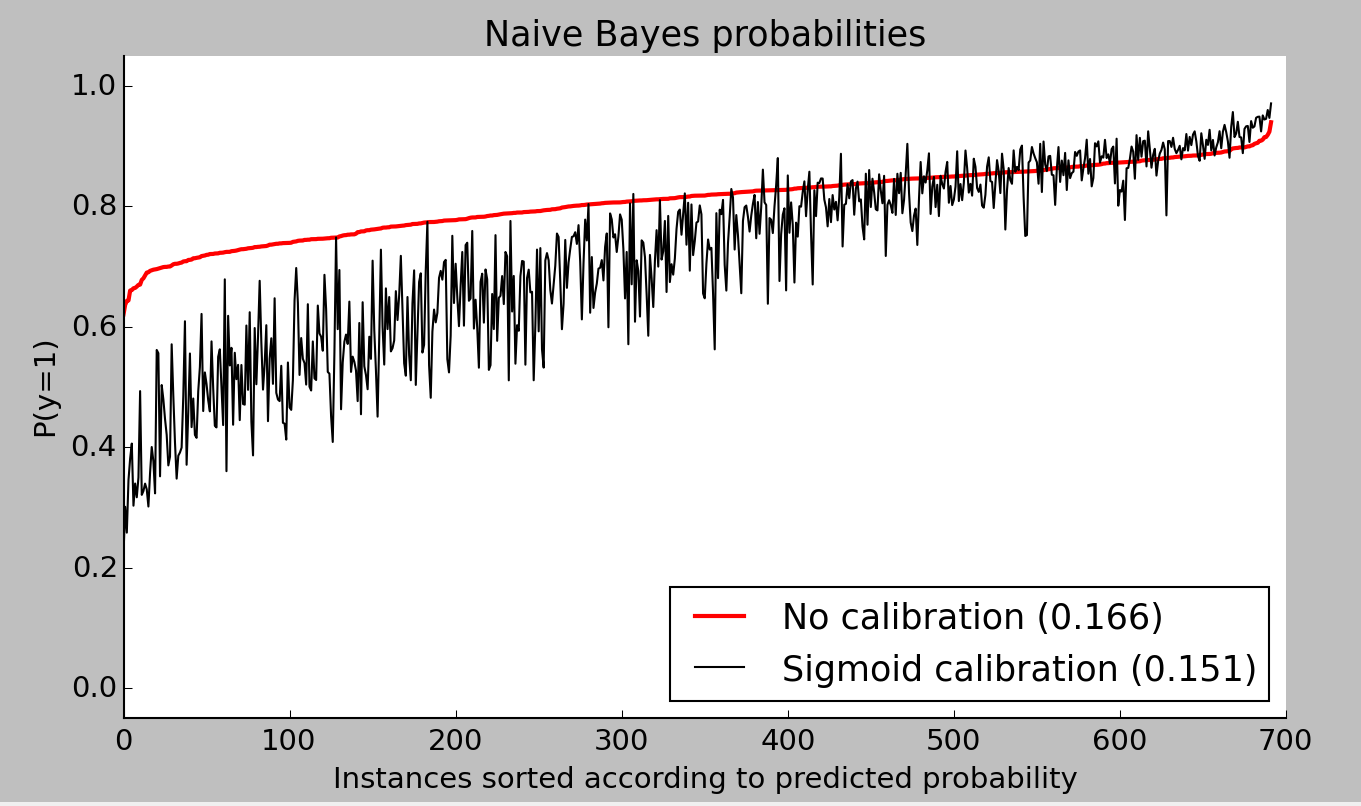


Figure-4 : Calibration results

As shown in the Figure-4, the sigmoid calibration results in a lower Brier score which indicates that the predictions are better calibrated. The Brier score measures the mean squared difference between (1) the predicted probability assigned to the possible outcomes for an item, and (2) the actual outcome. The figure-5 shows accuracy results of the model. The recall is intuitively the ability of the classifier to find all the positive samples. After calibration, the recall metric has gone down to 0.956 and at the same time precision metric has gone up from 0.75 to 0.80 reducing the false positive rates and improving the accuracy of prediction. The roc measure has gone down slightly from 0.83 to 0.78 but we are able to considerably improve the accuracy on test data reducing the possibility of over-fitting and under-fitting.

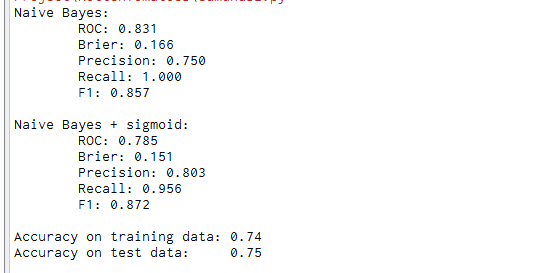


Figure-5 : Accuracy chart of the model

**Future Improvements**

The accuracy of the model on the test data is slightly better than the training data. So the benefits of further improvements may be minimal if all parameters in this study remain the same. So cross validation was not performed. To validate the classifier against unseen data, we can split the data in to k parts and then perform k-fold cross validation(Stewart, 2015). Then take the average score for each fold to get an accurate estimate of the performance. Other ways to improve the results include: fine tuning the Naive Bayes classifier, extract more features from the data, fine tune the ngrams etc.

**Conclusion**

Text mining techniques presented in this study are versatile and scalable across any industry where online consumer reviews play a role in market research. The unsupervised text analysis can be applied to extracting customer sentiments on higher granularity. The supervised text analysis can be applied to quantify the level of customer sentiments. These techniques find a greater application in social media where customer sentiment analysis play a greater role.

**References**

Grisel, O. (2010). Topics extraction with Non-Negative Matrix Factorization. Retrieved December 19, 2015, from http://scikit- learn.org/0.16/auto\_examples/applications/topics\_extraction\_with\_nmf.html

Notebook on nbviewer. (2010). Retrieved December 19, 2015, from http://nbviewer.ipython.org/github/cs109/content/blob/master/HW3\_solutions.ipynb

PR, N. (2014, April 9). Edmunds.com Mobile App, Live Advice Service Recognized by 18th Annual Webby Awards. PR Newswire US.

Stewart, Z.(2015, April 29). Retrieved December 19, 2015, from http://zacstewart.com/ 2015/04/28/document-classification-with-scikit-learn.html

WELCOME TO THE EDMUNDS API. (2015, August 25). Retrieved December 19, 2015, from http://developer.edmunds.com/