Analyzing Consumer Purchasing Behaviour

Project Code - P9
November 25, 2018

Supervisor -
Prof. Kiran Seth
Prof. Achal Bassamboo

Shubham Singla (2015ME10383)
Manas Joshi (2015ME10108)

INTRODUCTION

- **E-Retailing - An Easy & Informed way to Purchase Products**
  - Quality of Product - Allowing customers to make informed decisions
  - What kind of information will attract more customers
  - Lightning Deals - Medium to study Real time purchasing behaviour
- **Review Text and Ratings**
  - <u>Review Text</u> - I bought this for my husband who plays the piano. He is having a wonderful time playing these old hymns. (Rating - 5.0)
  - <u>Review Text</u> - I bought the product a month ago and it's stopped working. Have replaced batteries and yet it's stopped. The product while it was working is really good. I probably have got a defective item. (Rating - 3.0)

LITERATURE SURVEY

- Previous Research explored the impact of inventory information on consumer demand
- Ruomeng Cui, Dennis J. Zhang, and Achal Bassamboo[5] claim higher prior percentage claim information attracts more sales (Didn’t consider Review Text)
- Raymond Y.K Lau and Wenping Zhang, Peter D. Bruza[7] along with K.F. Wong - Determined the positivity and negativity of reviews
- Eun-Ju Lee and Soo Yun Shin[6] examined how the quality of online product reviews affects the participants’ acceptance of the reviews
- Reid Pryzant, Young-joo Chung , Dan Jurafsky[2] - Predict sales from product descriptions
- Anindya Ghose and Panagiotis G. Ipeirotis[4] try to predict the helpfulness rating given to a review from its text

- Deep neural networks have been found to perform better than conventional regression models across a number of different applications [13].
- Since the development of recurrent neural networks, or RNNs, these architectures have been heavily used to classify, process and predict natural language data.
- Also the introduction of long short-term memory [12], or LSTM, have transformed speech recognition and synthesis, machine translation, language modeling, and image captioning.
- James et al.[11] suggest skew in the dataset hampers the performance and the reviews in the central region are prone to noise, making the classification task at some boundary difficult.

OBJECTIVES

- Infer that review text along with rating gives more predictive information as compared each one individually.
  - Predicting helpfulness of a review from review text and rating combined
- Use the reviews, ratings and real time inventory availability to study the consumer purchasing behaviour for
  - Normal products
  - Products whose deal time is increased abruptly

METHODOLOGY

- To establish the significance of reviews, we require a way to extract useful information from them and an appropriate performance metric.
- We try to predict helpfulness ratio as our first task and establish the significance.
- Scrap the data of lightning deals from Amazon and study them
- Analyze consumer purchasing behaviour for normal and special products which have a jump in their deal time

THEORY

Linear Model

- Given an input vector $x \in R^m$, where $x_1, \ldots, x_m$ represent independent variables, a linear model tries to fit an equation of type $y = \beta_0 + x^T\beta$, linear in parameters β ,where $y \in R$ is a dependent variable.

Logistic Regression

- Logistic regression analogous to linear regression, measures the relationship between the categorical dependent variable and independent variables by estimating probabilities using a sigmoid function.

Random Forest Classifier

- Fits multiple Decision Tree Classifiers on Sub-samples of Dataset
- Based on building several estimators independently and averaging their predictions.
- On average, the combined estimator is usually better than any of the single base estimator because its variance is reduced.

Deep Neural Network

Word Embeddings

- Dense representation of english words as real valued numbers in a vector space
- Words with similar context having similar representations, naturally capturing their meaning.

Long Short Term Memory Network

- Order-insensitive models insufficient to fully capture the semantics of natural language.
- Recurrent neural networks designed to find pattern in sequential data, but fail to handle long range dependencies.
- LSTM networks are an extension to RNN, addressing problems of vanishing gradients by introducing a memory cell.

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right)$$
$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right)$$
$$c_t = f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right)$$
$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right)$$
$$h_t = o_t \tanh(c_t)$$

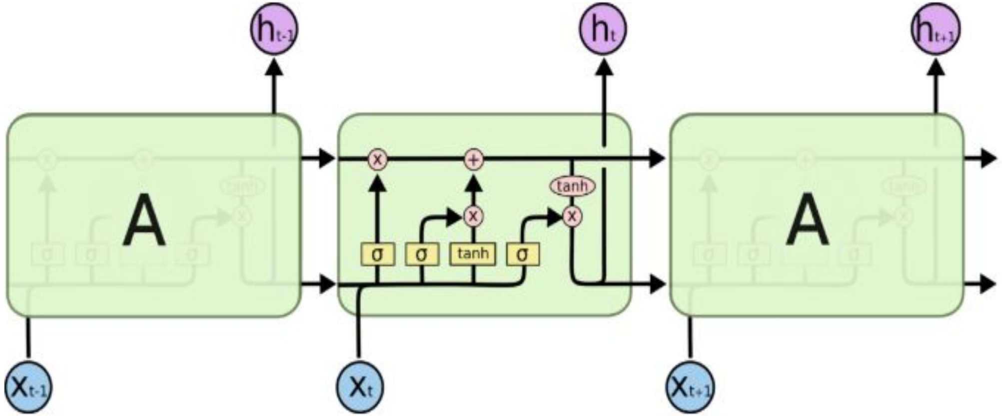


Fig 1 - LSTM Cell

DATA COLLECTION & FEATURE ENGINEERING

- Review dataset provided by Mcauley et al. [1]
  - Review text with rating and helpfulness votes
  - Product Metadata
- Model 1
  - Domain specific reviews with single classifier (0-1)
  - Traditional ML
    - Readability Features - Complexity of Review
    - Subjectivity Features - Dynamic Language Model with n-grams (n = 8)
  - Neural Network
    - Text as a feature to a deep neural network
- Model 2
  - Universal reviews with deskewing operation
  - Two classifiers, one for helpful and another for not helpful reviews
  - Text and metadata as a feature to a deep neural network.

Dataset Distribution

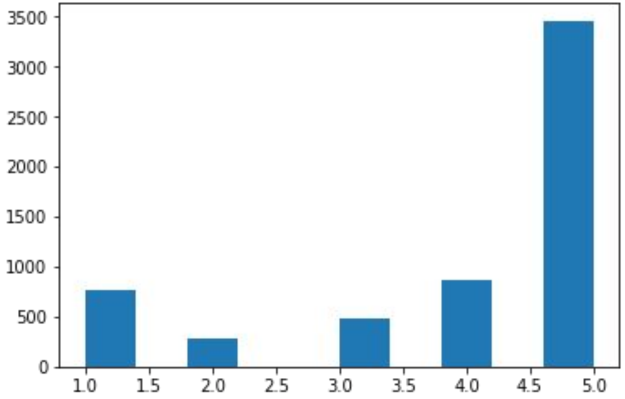


Fig 1 - Ratings Histogram

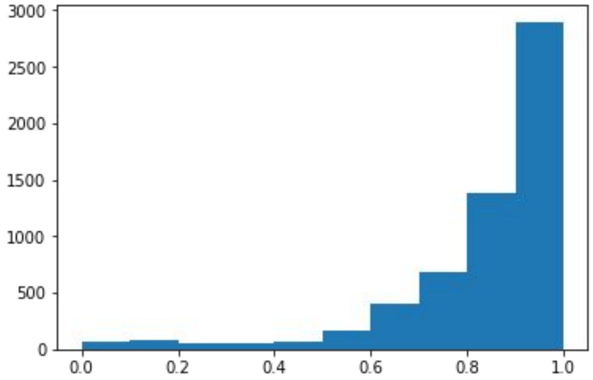


Fig 2 - Helpfulness ratio Histogram

Dataset	Not Helpful Reviews	Helpful Reviews
Train Data	8595	11068
Dev Data	2121	2795
Test Data	2693	3452

Table 5 - New Unskewed Dataset

RESULTS - Model 1

1. Linear Regression

   ○ Helpfulness = $\beta_0 + \beta_1$(Readability) + β_2(AvgProb) + β_3(DevProb)
   ○ R^2 Value increased from **0.06** (Review_text Only), **0.15** (Ratings only) to **0.20** (Combined)

2. Random Forest

Models	F1 Score (Class 0)	F1 Score (Class 1)
Random Forest on Review_Features	0.098	0.985
Logistic Regression on Ratings	0	0.961
Random Forest on Review_Features + Ratings	0.22	0.95

Table 3 - Helpfulness Ratio as a Binary Variable (Class 1)

3. Deep Neural Network

Model Used & Features Trained	F1 Score (Class 0)	F1 Score (Class 1)
DNN on reviews	0.608	0.67
Logistic Regression on Ratings	0.58	0.65

Table 6 - DNN Results for Class 1

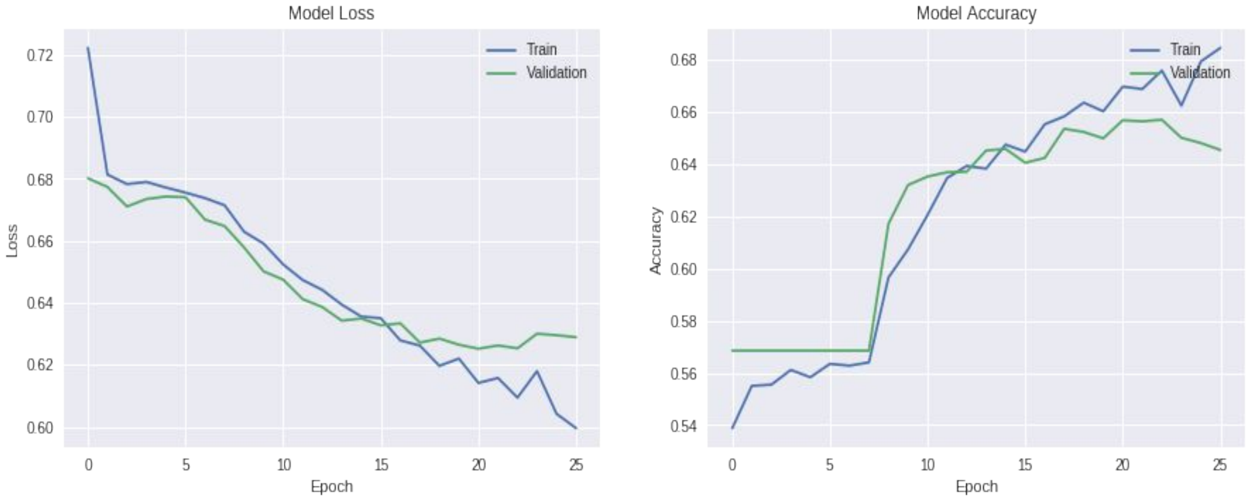


Fig 4 - Graph for Model loss (left) and Model accuracy (right) with each epoch

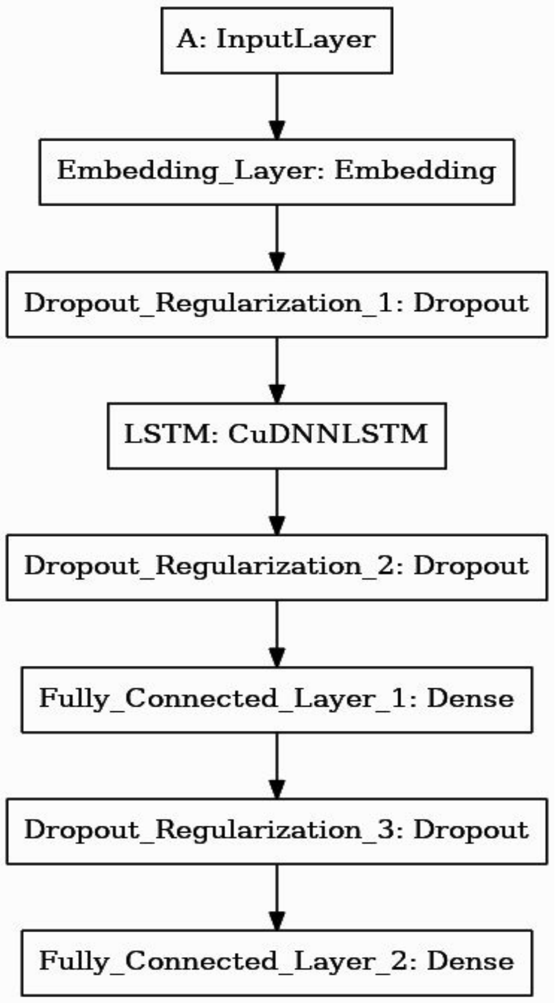


Fig 3 - DNN Layers

Improvement Attempts

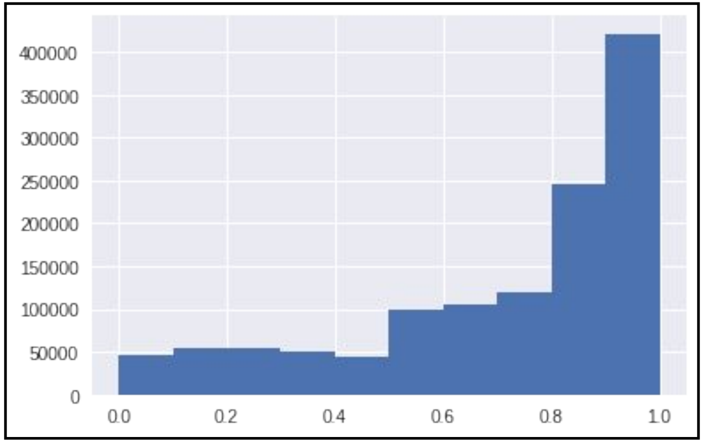
- 42552 out of vocabulary of 80474 not present in the Word2Vec model (spelling mistakes, brand names etc.)

- Use Fasttext Word Embeddings provided by Facebook Research.
    - Capable of forming vectors of unknown words by grouping n-grams of words.
    - No significant change, with similar precision and recall values. **Accuracy on Test data = 65%**

- Utilize a spell corrector to correct the spellings in the reviews
    - Poor performance with **accuracy 65% on Test data**

Model 2

- We choose the thresholds such that, a review is classified helpful if $f_r > 0.80$ and not helpful if $f_r \leq 0.20$
- A Deskewing step is performed to achieve an approximate uniform data.
- Two classifiers are trained, one for helpful and another for not helpful reviews



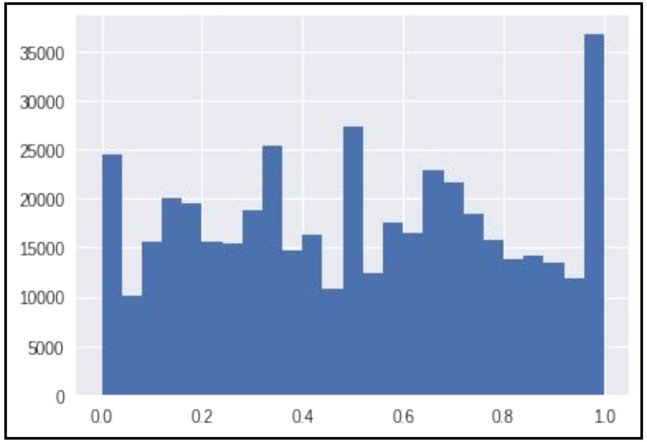


Figure 5 : Dataset distribution of helpfulness ratio: before deskewing(left) and after deskewing(right)

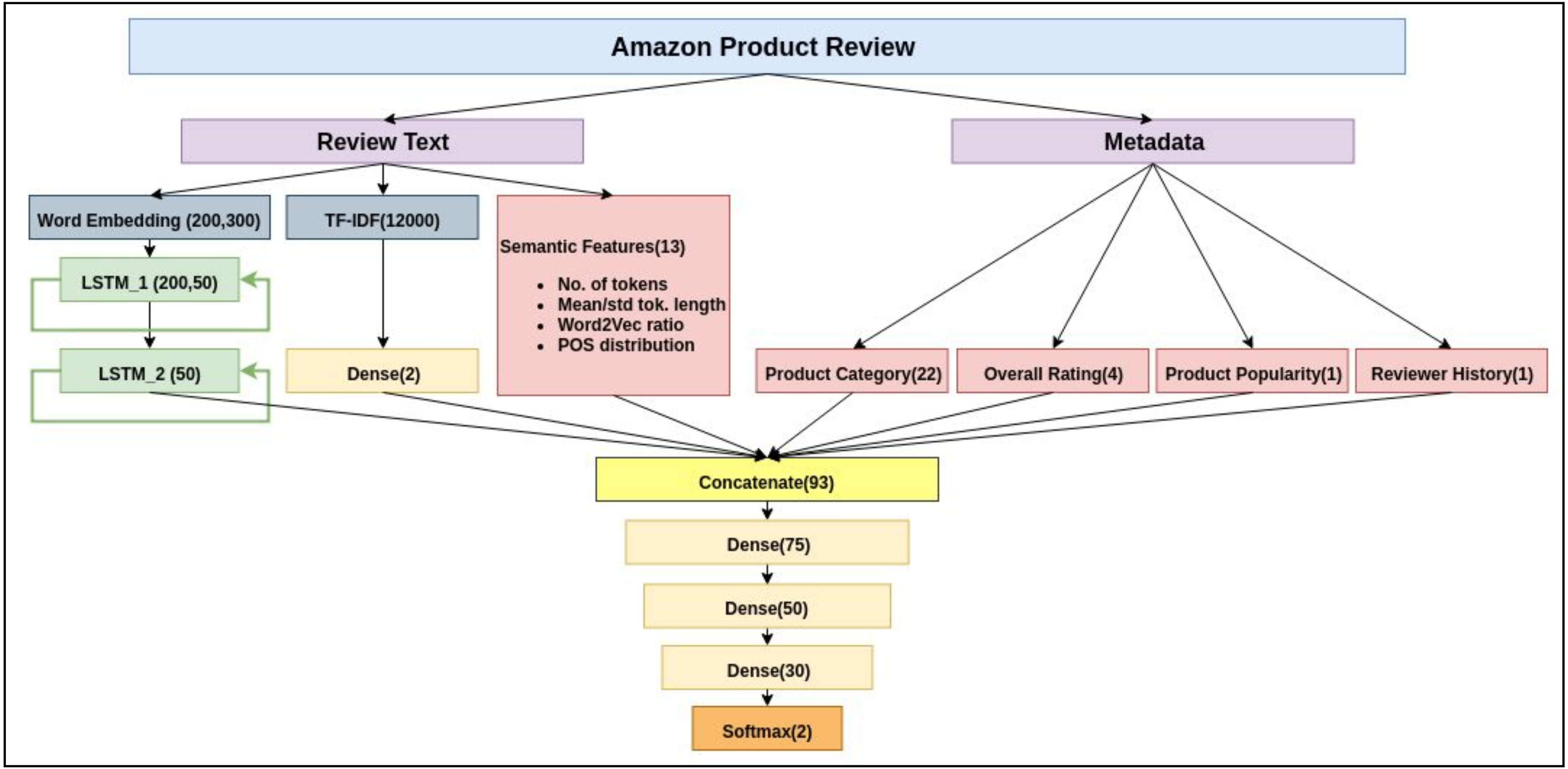
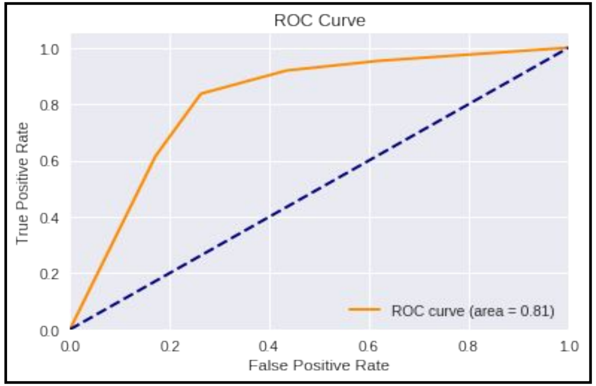


Figure 6: Architecture of Deep Neural Network

RESULTS

Results show us that this approach for analysing reviews is better compared to the previous one. It also favours our claim that the information in reviews and ratings combined is more than the ratings alone.

Features	**Test Accuracy**		**AUC**	
	Good Classifier	Bad Classifier	Good Classifier	Bad Classifier
Only Ratings	76.5	77.5	0.79	0.81
Reviews and Ratings	**80.5**	**83.7**	**0.88**	**0.90**



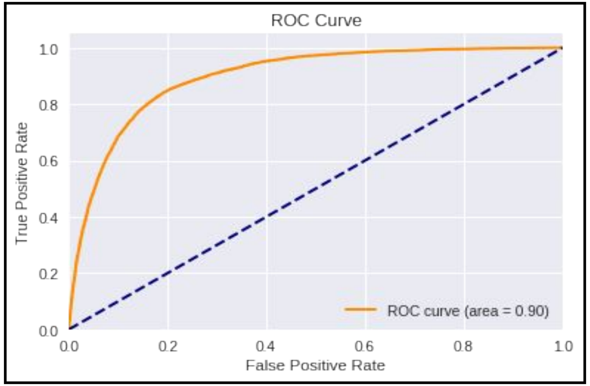


Figure 10 : ROC curve for bad classifier: ratings and review features(right) and rating features(left)

Problem 2 - Analyze consumer purchasing behaviour

- Data Scraping
  - Scrape per minute data of lightning deals : percentage of deals claimed and the time remaining.
  - Screen scraping
    - Slow and results in IP being blocked if high number of requests are sent.
    - Dependent on the Amazon's product placement on different pages.
  - Scrape by sending AJAX requests to Amazon server using XMLHttpRequest and Fetch Api.
  - Acquire per minute data for at least 300 products within 5 seconds using this.
- Data Statistics
  - Lightning Deals Collected - 1437 (November 4, 2018 - November 10, 2018)
  - Deals observing change in time - 847
  - Lightning Deals of Interest - 14
  - Observation time for Regression - 60 minutes
  - Total data points generated - 63467
  - Special Deal points - 164
  - Data points before time increase - 140
  - Data points after time increase - 24

Results

Regression Equation - $dealClaim_{i,t} = w_0 + w_1 \times C_{i,t} + w_2 \times numRev_{1t} + w_3 \times avgRating_{2t} + w_4 \times actualDisc_{3t} + w_5 \times dealDisc_{4t} + w_6 \times dummy_{i,t} + w_7 \times dummy_{i,t} \times remDealTime_{rem} + w_8 \times remDealTiime_{rem}$

Variables	Coefficient	Std error	t-value	P > | t-value |
$x_0 = 1$	-7.153	8.685	-0.824	0.411
$C_{i,t}$	-0.0616	0.033	-1.885	0.061
$numRev_{2t}$	0.0012	0.001	2.157	0.033
$avgRating_{3t}$	1.5437	1.753	0.881	0.380
$actualDisc_{4t}$	0.0358	0.108	0.332	0.740
$dealDisc_{5t}$	0.0939	0.126	0.747	0.456
$dummy_{6t}$	2.6091	3.9	0.669	0.504
$dummy*remDealTime_{7t}$	-2.85×10^{-7}	1.43×10^{-7}	-0.2	0.842
$remDealTime_{8t}$	-1.05×10^{-8}	3.98×10^{-8}	-2.641	0.009

1. P value for dummy variable and interaction term is not significant.

2. We need more data of special deals

3. Present dataset is skewed and contains very less data points

CONCLUSION AND FUTURE WORK

- Review text along with review rating is a better metric as compared individually.
- Analyzing how customer purchasing behaviour depends upon real time inventory
- Study the data manually to look out for any abrupt changes in product data
- Analyzing consumer purchasing behaviour for deals having increase in deal time
- Analyzing consumer purchasing behaviour for normal lightning deals
- Modelling the sales as a function of review text, ratings and other features

REFERENCES

[1] Ups and downs: Modeling visual evolution of fashion trends with one-class collaborative filtering, R. He, J. McAuley, *WWW*, 2016

[2] W.H. DuBay, The Principles of Readability, Impact Information, http://www.nald.ca/library/research/readab/readab.pdf, 2004.

[3] B. Pang and L. Lee, “A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts,” Proc. 42nd Ann. Meeting of the Assoc. for Computational Linguistics (ACL ’04), pp. 271-278, 2004.

[4] Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics Anindya Ghose and Panagiotis G. Ipeirotis, IEEE, 2011

[5] Learning from Inventory Availability Information: Evidence from Field Experiments on Amazon; Ruomeng Cui, Dennis Zhang, Achal Bassamboo, Forthcoming Management Science, 2016, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2868218

[6] When do consumers buy online product reviews? Effects of review quality, product type and reviewer’s photo; Eun Jee, Soo Yun Shin IEEE 2014

[7] Learning Domain-Specific Sentiment Lexicons for Predicting Product Sales; Raymond Y.K Lau, Wenping Zhang, Peter D. Bruza, KF Wong, IEEE 2011

[8] Alias-i. 2008. LingPipe 4.1.0. http://alias-i.com/lingpipe (accessed October 1, 2008)

THANK YOU