Prediction of Rises and Falls of Business Ratings based on early YELP User's Reviews

R Capstone Project, Victor M

November 18, 2015

1. Introduction

Yelp database includes dozens of parameters related to businesses as well as their customers. One of the most important indexes for a business is the rating from its customers. The present research is devoted to the rating of a business. In particular we analyze which factors from the Yelp Dataset accompany businesses with high and low ratings.
Occasionally ratings could change in time. Luckily the database makes it possible to track the changes and pick out groups of businesses with positive and negative trends. Our idea that the trend can be predicted based on the information gathered during relatively short initial period. Basically this research could help business owners to see how they can benefit from the information provided by the Yelp review service.

1.1 The Yelp Dataset

The version of the Yelp Academic Dataset used in the analysis was released by Yelp on August 2015. The dataset consists of five main objects encoded as JSON files:

- businesses
- checkins
- reviews
- tips
- users

The recorded information spans a more than ten year period, with the earliest data from October 2004, and the most recent data from January 8, 2015. The scope of our study utilized all the objects, which provide over 1,500,000 reviews from more than 60,000 businesses.

Important identifiers considered include business ids, ratings, dates, reviews, likes, tips, votes etc.

1.2 Pre Processing of the Data

We read JSON format and convert it into R objects Data Frame. The relevant files are stored in "data" folder.

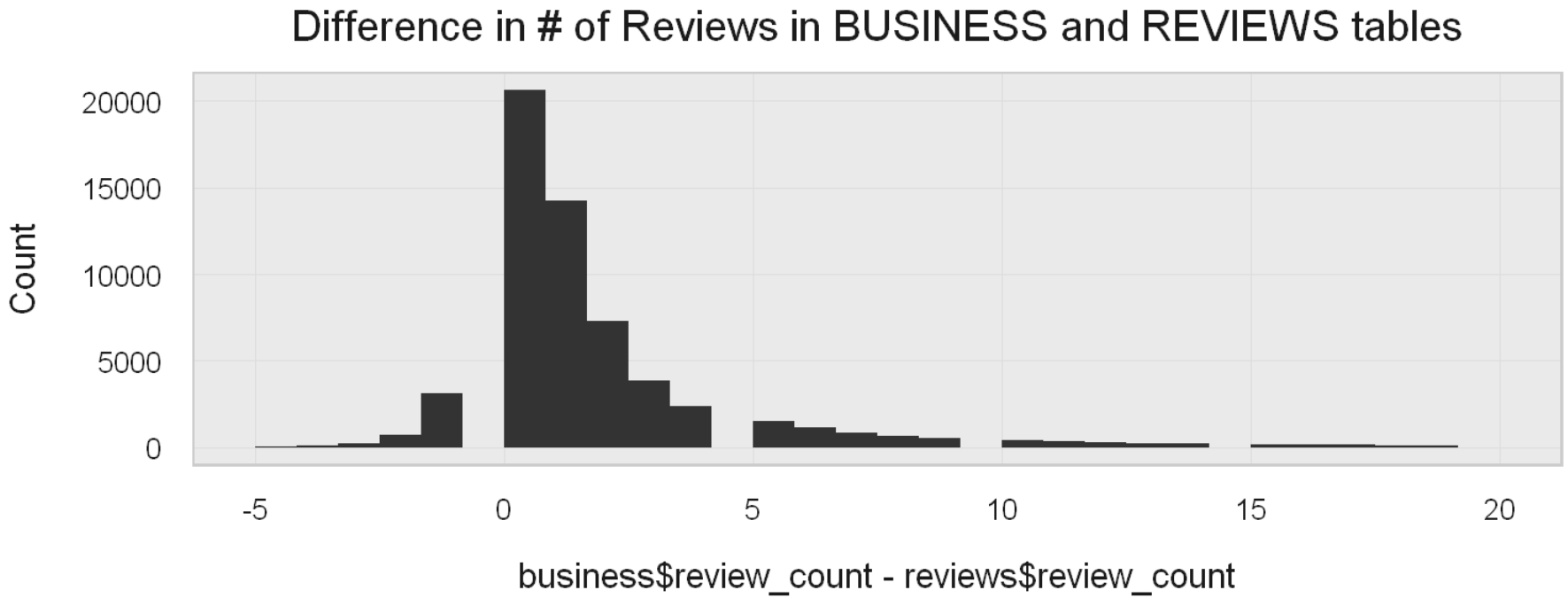
2. Methods

2.1 General Information

We use standard methods of exploratory data analysis to make a general representation of rating distribution over all the businesses registered in the Yelp Database. We find dependencies between rating and characteristics of the submitted reviews.
Additionally we will check if the distribution of star rating is identical for the highly reviewed businesses and businesses with a small number of reviews.
We develop three predictive models which make it possible to say whether the rating of a business will increase or decrease after a while. The forecast is solely based on the information gathered during relatively short period of time (we set this period as time needed for a business to receive the first 5 reviews).

2.2 Challenges with the Data

An inconsistency in the dataset was found out: the number of reviews in BUSINESS table differs from that in REVIEW table. The following histogram shows the distribution of the difference.



Notes: 1. NAs in review_count from the REVIEW table were converted into zeros. 2. Extreme differences from both sides were excluded from the plot to show the shifted distribution.
The statistical parameter of the distribution are following:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## -27.000   0.000   1.000   2.614   2.000 476.000     399
```

It can be seen that the BUSINESS table accounts for much more reviews than submitted in the REVIEW table. Despite of this fact we will use mainly data from the REVIEWS table, as it contains detailed information on every review for every business.

Similar situation can be found for the rating of businesses. In some cases average star are not equal in these two tables. There are **3565 businesses out of 61184 with the difference 0.5 and more**.
These facts show that the Dataset is probably incomplete or has been somehow filtered. Thus we will further rely on the information provided in REVIEWS tables.

2.3 Rises and Falls: definition

Business ratings could change in time. The only factors within the Dataset that can be tracked in time are the review text, star, tip text. Unfortunately the rest of factors are not recorded chronologically.
We pick out groups of businesses with positive and negative trends in order to track the changes in ratings. Our idea that the trend can be predicted based on the information gathered during relatively short initial period (5 reviews).

2.4 Aggregation of Relevant Data

At first we define **the initial rating** as the mean rating **collected from the first 5 reviews**. For this period we additionally calculate the following parameters:

- coefficient of variation (it shows the variability of rating)
- average number of days needed for a business to gain review
- mean text length of the reviews
- mean text length of the tips

We created subgroups of businesses based on the average rating change:

- inc (increasing)
- dec (decreasing)

This classification is stored in the ‘group’ variable.

Sample of the Aggregated Business Data

business_id	r_count	r_stars_avg	r_length_avg	r_stars_1_avg	r_stars_1_cv	r_length_1_avg	r_time_1_avg	r_stars_diff	group
vcNAWiLM4dR7D2nwwJ7nCA	10	3.60	543	3.8	0.29	447	456	-0.20	dec
mVHrayjG3uZ_RLHkLj-AMg	10	4.70	774	4.8	0.09	583	85	-0.10	dec
KayYbHCt-RkbGcPdGOThNg	12	3.92	628	3.8	0.12	846	232	0.12	inc
b9WZJp5L1RZr4F1nxclOoQ	34	4.62	606	4.2	0.20	474	178	0.42	inc

P1fJb2WQ1mXoiudj8UE44w	45	3.49	700	4.6	0.12	702	230	-1.11	dec
3gmBc0qN_LtGbZAjTtHWZg	16	3.38	427	3.0	0.41	376	175	0.38	inc

Note 1: *businesses with at least 10 reviews are considered (25977 out of 61184 businesses)*
Note 2: *the description of the variables (column names) could be found in the section 3.4*

In the following chapter we will develop three models which predict in which group a business falls depending on the initial period parameters.

2.5 Parameters of the predictive modelling process

First, we split the data into two groups: a training set and a test set. The percentage of the data in the training set is accepted 70%.

2.6 Predictive models

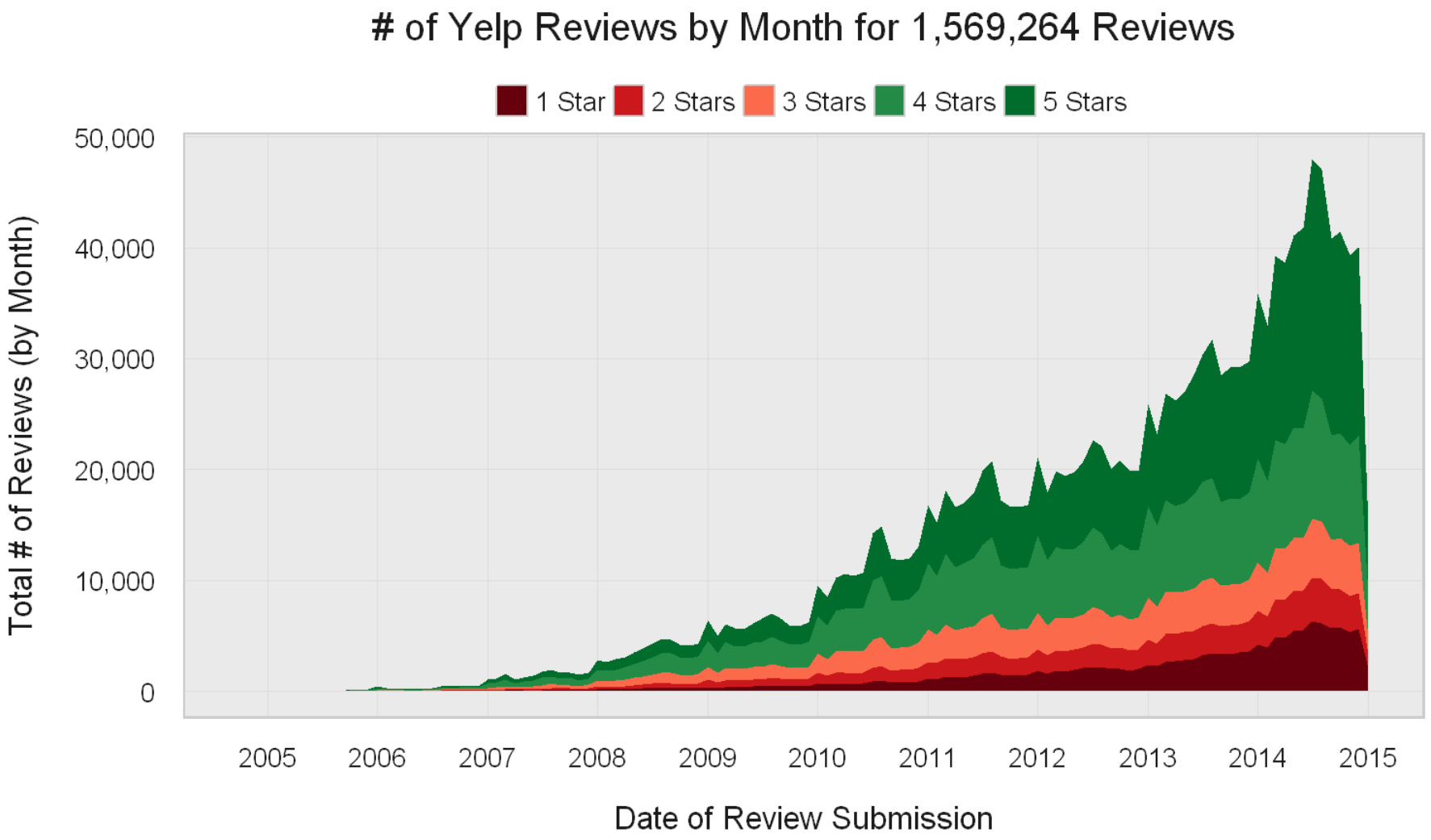
Three decision models were created with RвЂ™s caret package based on the above mentioned predictor variables. We use the following models to predict the two classes:

- **pls** - Partial Least Square Discriminant Analysis
- **glm** - Generalized Linear Model
- **knn** - k-Nearest Neighbors

3. Results

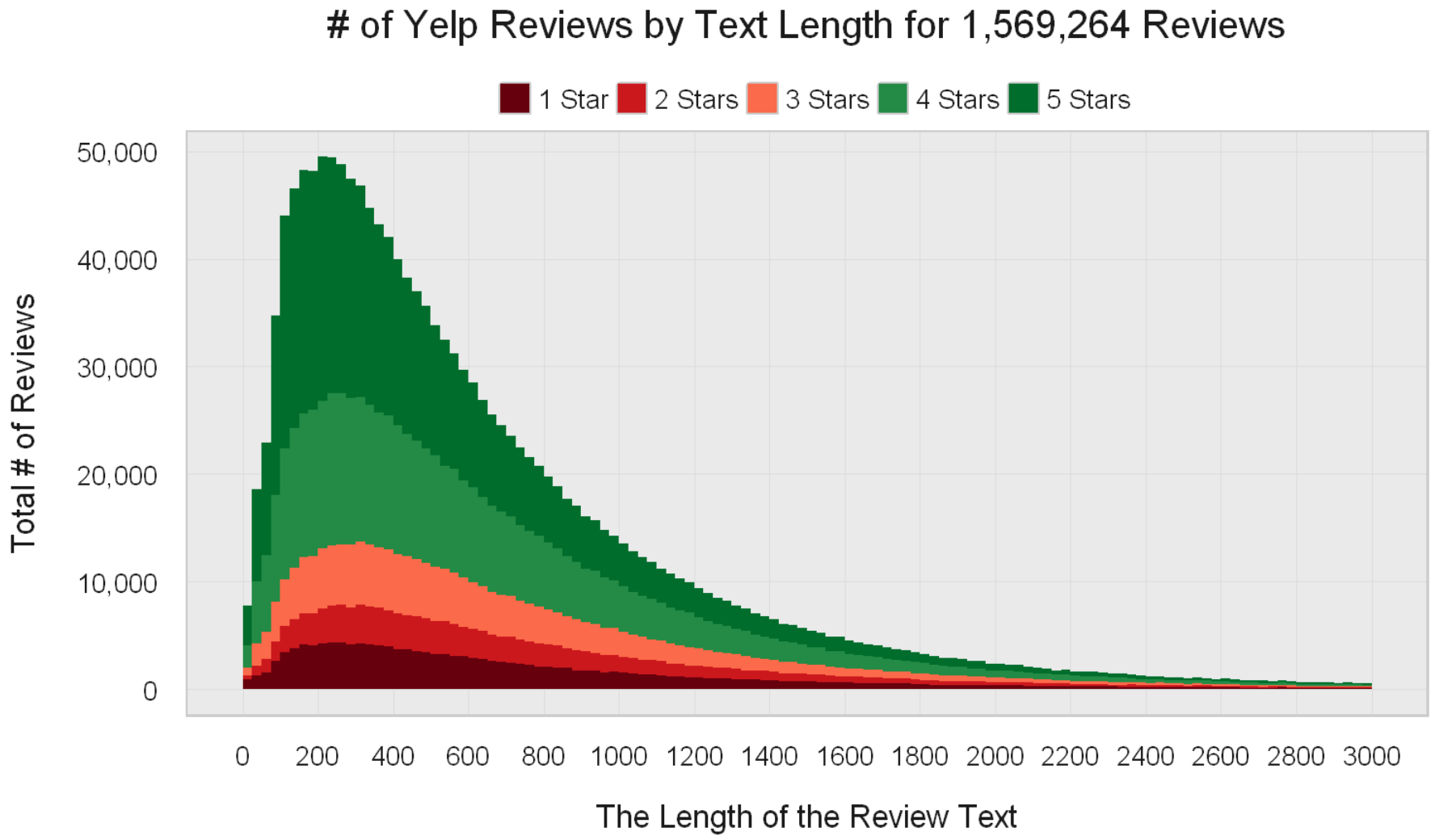
3.1 Ratings Overview

The following histogram illustrates review accumulation over time for all the businesses taking into account the star rating. It demonstrates the incredible growth in the number of new reviews since 2005. We can see that the proportion of 1-2-3-4-5-stars submitted by users remain approximately equal.

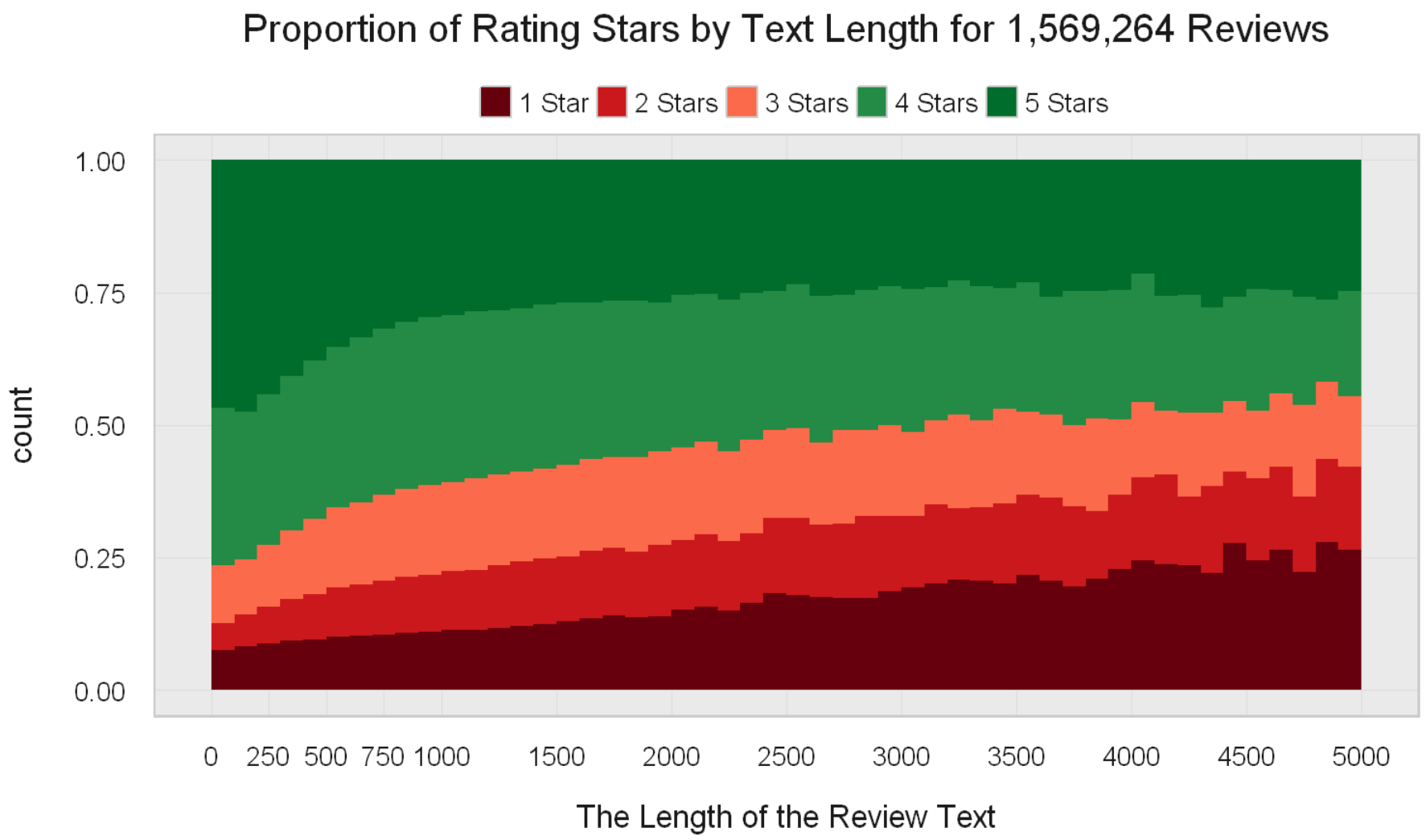


3.2 Influence of the Review Text Length

The reviews can be characterized by text length. The following histogram shows distribution of reviews by number of symbols in review text. The histogram is truncated on 3000 symbols as the number of reviews with a longer text is relatively small.



The following diagram makes it possible to see how the rating and text length are related.

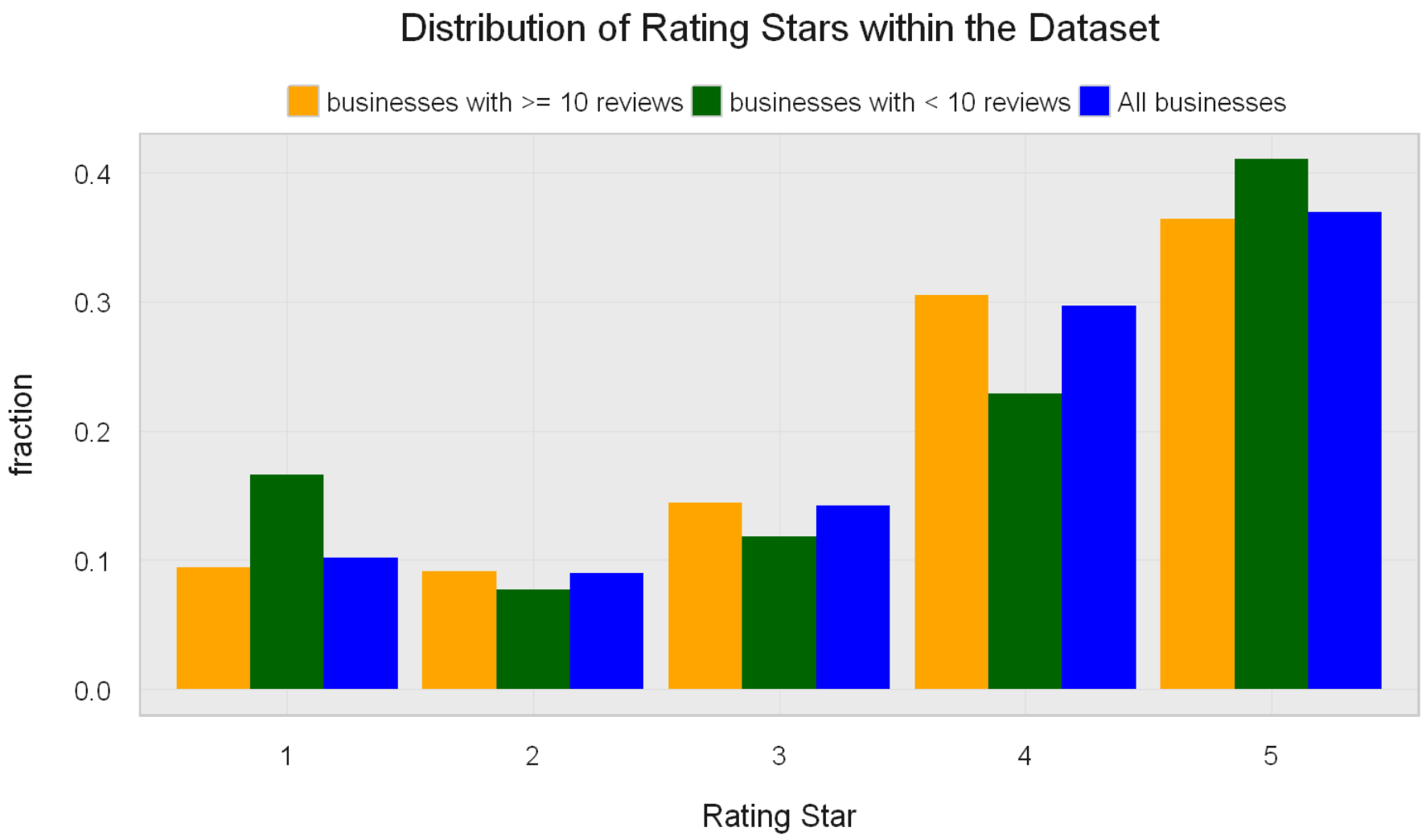


It can be seen that shorter reviews are more likely positive. For instance a short review (less than 200 symbols) is positive (4 or 5 stars) with probability at least 75%. On the other side, for long reviews (2500 symbols and more) the probability of being positive falls to about 50%. This fact will be exploited further.

3.3 Distribution of Star Ratings over all the Reviews

The following image represents the distribution of 1 to 5 stars over all the submitted reviews (blue color). The diagram makes it possible to reveal the difference in stars distribution for highly reviewed businesses (orange color) and for businesses with a pretty small number of reviews (green color).
We divided all the businesses into two subgroups depending on the corresponding review count: businesses with at least 10 reviews submitted, and businesses with less than 10 reviews. Then for each subgroup the distribution of star ratings was calculated. The results are reflected on

the plot.



It should be noted that the distribution of star rating for highly reviewed businesses (orange color) is similar to the overall distribution (blue color); while the distribution of ratings for lowly reviewed businesses (green color) has heavier tails. The similarity of the overall distribution and filtered distribution for highly reviewed businesses proves that we can safely skip lowly reviewed businesses from the analysis.
An explicit bimodal form of the distributions can be explained psychologically: people tend to write reviews and rate places mostly after having an extreme experience, good or bad.

3.4 Predictive modelling

3.4.1 Working Data

Column Name	*Description*
business_id	Business identification
r_stars_avg	Average star rating for a business
r_stars_1_avg	Average stars over the 1st period
r_stars_1_cv	Coefficient of variation of stars over the 1st period
r_length_1_avg	Average review text length over the 1st period
r_time_1_avg	Average time interval between consequent reviews over the 1st period
r_stars_diff	Difference of ratings between the 1st period and overall rating
group	The category of a business according to the changes of its star rating:
	dec (decreasing, r_stars_avg - r_stars_1_avg < 0);
	inc (increasing, r_stars_avg - r_stars_1_avg >=0)

3.4.2 Dependent Variable

The variable **group** is considered as dependent. It is a factor variable and it implies the classification type of problem.

3.4.3 Independent Variables

The following variables are considered as dependent (predictors):

- **r_stars_1_avg**
- **r_stars_1_cv**
- **r_time_1_avg**
- **r_length_1_avg**

3.4.4 Prediction

First, we split the data into two groups: a training set and a test set. The percentage of the data in the training set is accepted 70%. we use the following models to predict the two classes:
* **pls** - Partial Least Square Discriminant Analysis
* **glm** - Generalized Linear Model
* **knn** - k-Nearest Neighbors

Performance parameters of the models such as ROC (reliability operational curve), Sensitivity, and Specificity are calculated using 30 random resamples for each model. The results of training three separate models are the following.

```
## $ROC
```

```
##       Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
## pls 0.7749  0.7842 0.7922 0.7911  0.7967 0.8161    0
## glm 0.7716  0.7836 0.7916 0.7912  0.8002 0.8063    0
## knn 0.7408  0.7470 0.7562 0.7560  0.7623 0.7820    0
##
## $Sens
##       Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
## pls 0.8510  0.8588 0.8665 0.8665  0.8725 0.8836    0
## glm 0.8466  0.8528 0.8651 0.8635  0.8720 0.8888    0
## knn 0.7855  0.8002 0.8043 0.8055  0.8110 0.8260    0
##
## $Spec
##       Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
## pls 0.4590  0.4768 0.4939 0.4922  0.5057 0.5380    0
## glm 0.4468  0.4821 0.5023 0.5006  0.5152 0.5578    0
## knn 0.5076  0.5323 0.5410 0.5399  0.5467 0.5684    0
```

After all the models have been trained, the confusion matrices were computed. Associated statistics for the models fit are shown below.

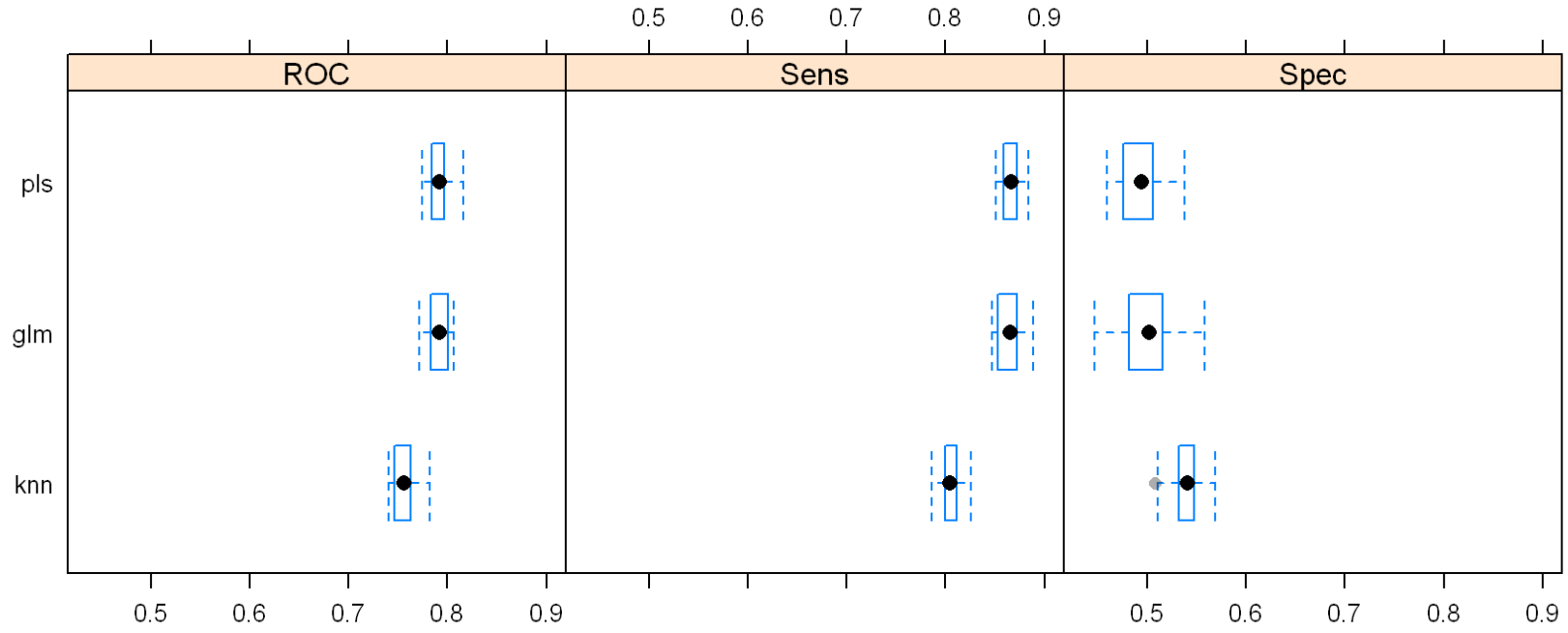
Performance Parameters for the Models

	Accuracy	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value
pls	0.7329655	0.8727126	0.4865248	0.7498272	0.6842893
glm	0.7336071	0.8688920	0.4950355	0.7521323	0.6816406
knn	0.7070448	0.8071587	0.5304965	0.7519670	0.6093686

Note: the 'dec' (decrease) class is considered as positive

4. Discussion

The results related to all the above mentioned predictive models looks similar except 'knn' model (see the plot).



A paired t-test can be used to assess whether there is a difference in the average resampled area under the ROC curve for the models.

```
##
## Call:
## summary.diff.resamples(object = diffs)
##
## p-value adjustment: bonferroni
## Upper diagonal: estimates of the difference
## Lower diagonal: p-value for H0: difference = 0
##
## ROC
##     pls       glm        knn
## pls           -7.349e-05  3.511e-02
## glm 1                     3.518e-02
## knn 1.801e-14 < 2.2e-16
##
## Sens
##     pls    glm      knn
## pls        0.002987 0.061013
## glm 0.6892          0.058026
```

```
## knn <2e-16 <2e-16
##
## Spec
##     pls       glm       knn
## pls           -0.008308 -0.047619
## glm 0.3183              -0.039311
## knn 1.169e-09 1.416e-08
```

Based on this analysis, the difference between the models is -7.349e-05 ROC units (the pls-model is slightly higher) and two-sided p-value for this difference is 1. The very low value of the difference shows that ‘glm’ and ‘pls’ models give almost identical results and can be both used.

It should be noted that the developed models are not perfect. They predict about 75% cases of rating decrease and about 67% cases of rating increase. These predictions can be helpful for businesses, because the future changes can be foreseen just using the first 5 reviews submitted by users.

Interestingly, the results differ significantly in case we change the initial number of reviews from 5 to 2. In accord with the calculations the models predict about 81% cases of rating decrease and about 71% cases of rating increase. However completeness of information from only 2 reviews is pretty small, so we recommend using first 4 or 5 reviews.