Project discussion:

Dataset on Blog Feedback

Tetiana Zverieva
Diana Kryskuv
Olha Soliar

14 June 2019

Summary

Dataset & Data Cleaning

- Data originates from **blog posts**
- The prediction task associated with the data is the prediction of the **number of comments** in the upcoming 24 hours
- Original dataset contains one **train** set and multiple **test** sets. Combining all test files into one is **needed**
- **Dimension**:
  - ***281*** attributes(***280*** features and ***1*** target variable);
  - ***52397*** individuals in train data;
  - ***7264*** individuals in test data.

  A **large** set of features.
- Dependent variable is highly **skewed**.
  - ***64.05%*** of it are zero;
  - among ***75%*** are less than ***10;***
  - can be high as ***1424***;
- Cleaning of the data is **not needed**

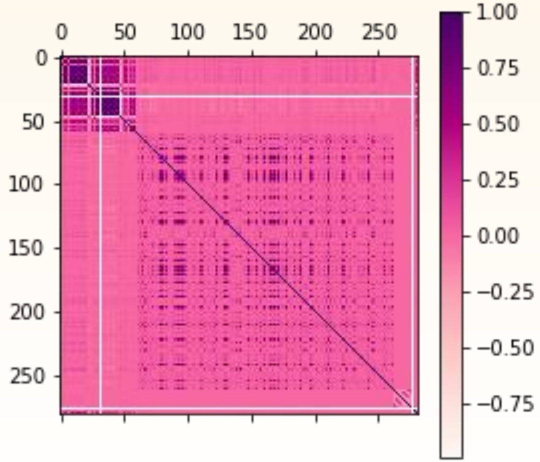
Description

Attribute Information:

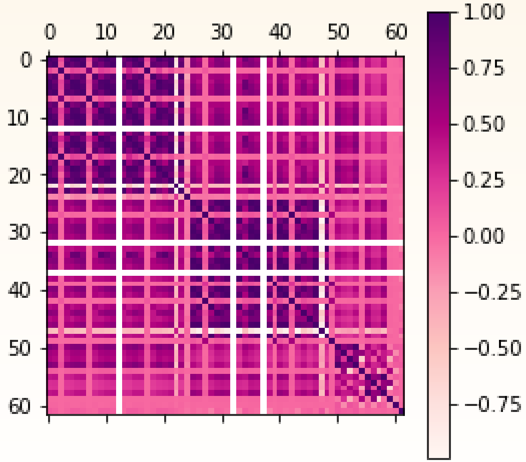
- *1...61:* Information about comments to a blog post:
  - number of comments and links in different periods;
  - average, standart deviation, min, max and median of number of comments.
- *62:* The length of the blog post.
- *63...262:* The 200 bag of words features for 200 frequent words of the text of the blog post.
- *263...276:* binary indicator features (0 or 1) for the weekday (Monday...Sunday) of the basetime and day of publication of the post.
- *277...280:* Information about parent blog post:
  - number of pages ;
  - minimum, maximum, average number of comments that the parents received.
- *281:* The target: the number of comments in the next 24 hours

Exploratory data analysis

- Representation of **correlation matrix**



- Correlation matrix of **highly correlated** values

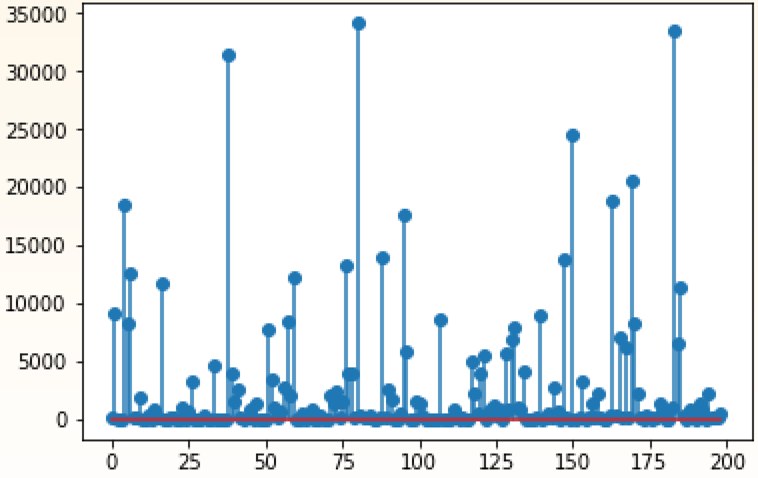


Exploratory data analysis

- Most **correlated** attributes with dependent variable

Nº	***Description of the feature***	***Correlation value***
9	median of number of comments in the last 24h before the basetime	0.506540
20	average of the difference between number of comments during 24h	0.503375
5	average of number of comments in the last 24h before the basetime	0.497631
4	median of the total number of coments before basetime	0.491707

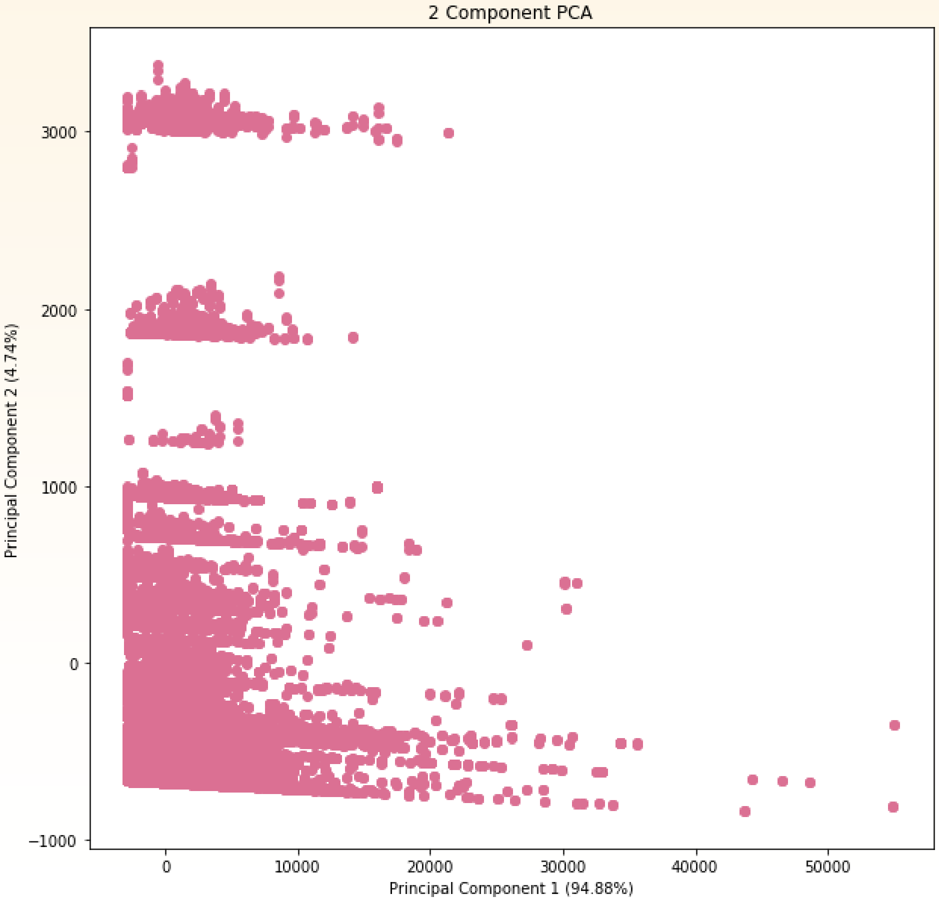
Exploratory data analysis



- ***200*** independent variables for the **words** contained in blog post;
- ***3*** words appear in more than **half** of the train data (probably stop words)
- ***117*** words appear in less than ***523*** observations ≈ ***1%*** of observations

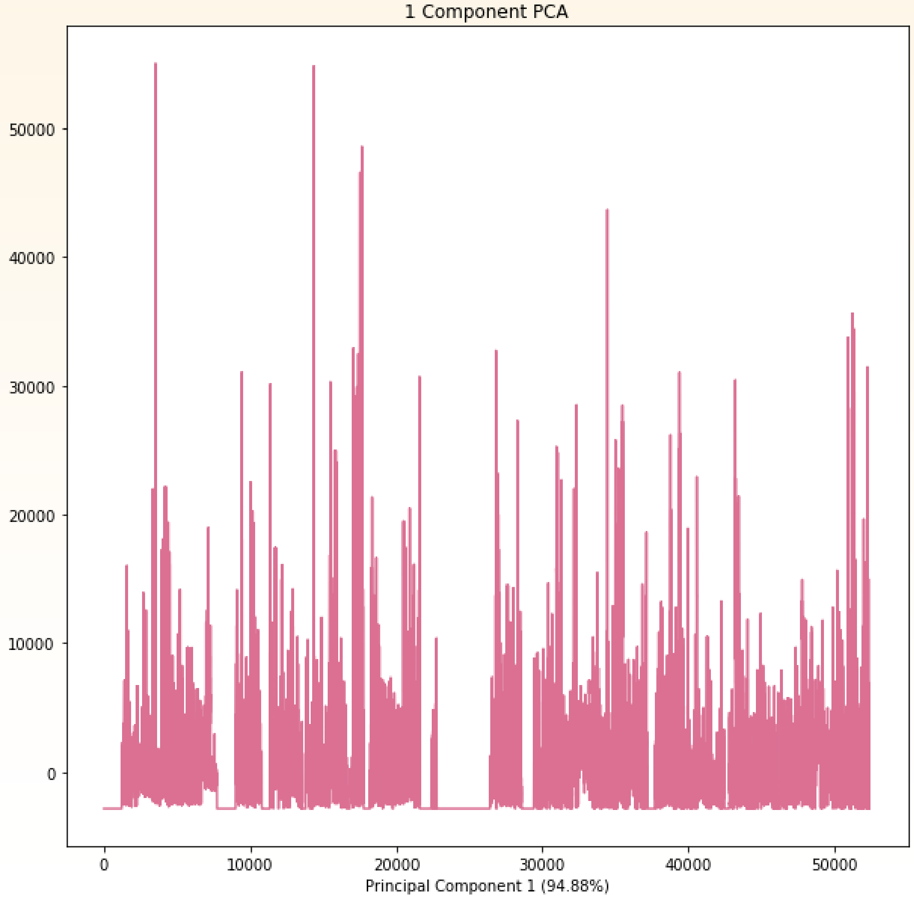
Principal Component Analysis

- Optimal number of components for representing 95% of data is 2
- Explained variance ratio of the first component is 0.9487
- Explained variance ratio of the second component is 0.0474



Principal Component Analysis

- Even with first variable we can explain almost 95% of the variance:



Linear Models Analysis(training dataset)

Ridge model results:

- Features selected by algorithm : 276
- R^2 value is 0.3591
- MSE of the best model is 911.06

Lasso model results:

- Features selected by algorithm : 42
- R^2 value is 0.3595
- MSE of the best model is 910.62

Non-Linear Model Analysis(training dataset)

Regression trees model results:

- Features selected by algorithm : 280
- R^2 value is 0.9859
- MSE is 19.946

Models Results (test dataset)

method	*Feature selected*	R^2	*MSE*
Lasso	42	0.3145	637.66
Ridge	276	0.3135	638.88
Regression trees	280	-0.19877	1115.15

- *In regression trees model we get an overfitting*

Models Results (test dataset)

- From results we can see that regression trees model performs very bad and we even get negative R^2 value, which means that our regression line is worse than using the mean value.

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$ESS = \sum(\hat{y}_i - \bar{y})^2$$
$$TSS = \sum(y_i - \bar{y})^2$$
$$RSS = \sum(y_i - \hat{y}_i)^2$$

Conclusions

- Regression tree model is overfitted
- Ridge and Lasso techniques are a great alternative when we are dealing with a large set of features
- For our dataset Lasso works well, because it shrinks the less important feature’s coefficient to zero thus, removing some feature altogether

Thanks for attention!