Muhammet Kara 171805036

Tuğçe Çördük 171805006

Furkan Gümrükçü 171805057

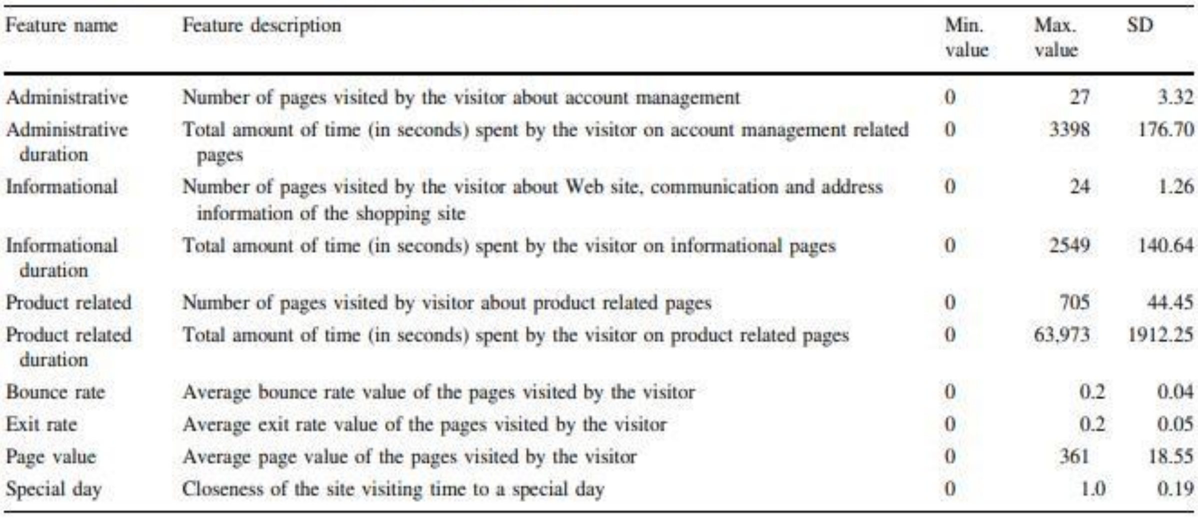
Classification Dataset:

Online Shoppers Purchasing Intention Dataset Data Set

The dataset consists of 10 numerical and 8 categorical attributes. The 'Revenue' attribute can be used as the class label.

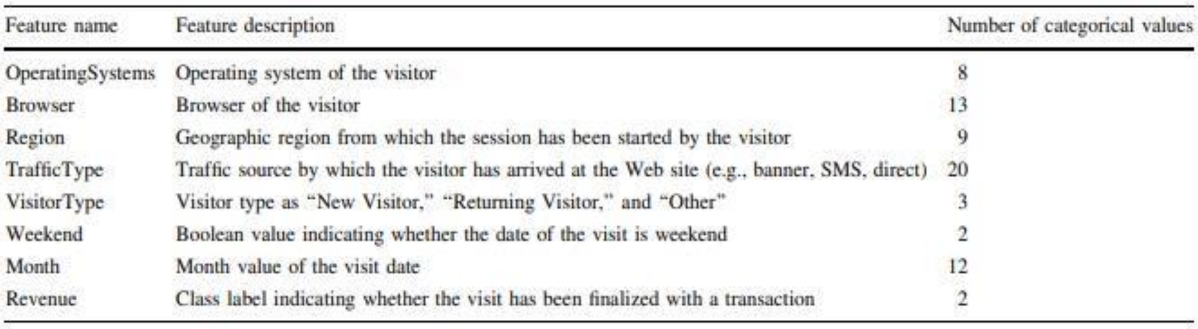
The numerical and categorical features will be used in the purchasing intention prediction model are shown in Tables 1 and 2, respectively. The dataset consists of feature vectors belonging to 12,330 sessions. The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period. Of the 12,330 sessions in the dataset, 84.5% (10,422) were negative class samples that did not end with shopping, and the rest (1908) were positive class samples ending with shopping.

Table 1 Numerical features used in the user behavior analysis model



Feature name	Feature description	Min. value	Max. value	SD
Administrative	Number of pages visited by the visitor about account management	0	27	3.32
Administrative duration	Total amount of time (in seconds) spent by the visitor on account management related pages	0	3398	176.70
Informational	Number of pages visited by the visitor about Web site, communication and address information of the shopping site	0	24	1.26
Informational duration	Total amount of time (in seconds) spent by the visitor on informational pages	0	2549	140.64
Product related	Number of pages visited by visitor about product related pages	0	705	44.45
Product related duration	Total amount of time (in seconds) spent by the visitor on product related pages	0	63,973	1912.25
Bounce rate	Average bounce rate value of the pages visited by the visitor	0	0.2	0.04
Exit rate	Average exit rate value of the pages visited by the visitor	0	0.2	0.05
Page value	Average page value of the pages visited by the visitor	0	361	18.55
Special day	Closeness of the site visiting time to a special day	0	1.0	0.19

Table 2 Categorical features used in the user behavior analysis model



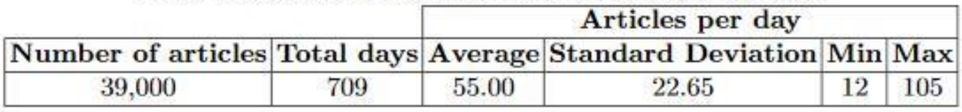
Feature name	Feature description	Number of categorical values
OperatingSystems	Operating system of the visitor	8
Browser	Browser of the visitor	13
Region	Geographic region from which the session has been started by the visitor	9
TrafficType	Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct)	20
VisitorType	Visitor type as "New Visitor," "Returning Visitor," and "Other"	3
Weekend	Boolean value indicating whether the date of the visit is weekend	2
Month	Month value of the visit date	12
Revenue	Class label indicating whether the visit has been finalized with a transaction	2

Regression Dataset:

Online News Popularity Data Set

The data retrieved the content of all the articles published in the Mashable, which is one of the largest news websites. The data was collected during a two year period, from January 7 2013 to January 7 2015. Very recent articles (less than 3 weeks) are discarded, since the number of Mashable shares did not reach convergence for some of these articles (e.g., with less than 4 days). After such preprocessing, dataset ended with a total of 39,000 articles, as shown in Table 1.

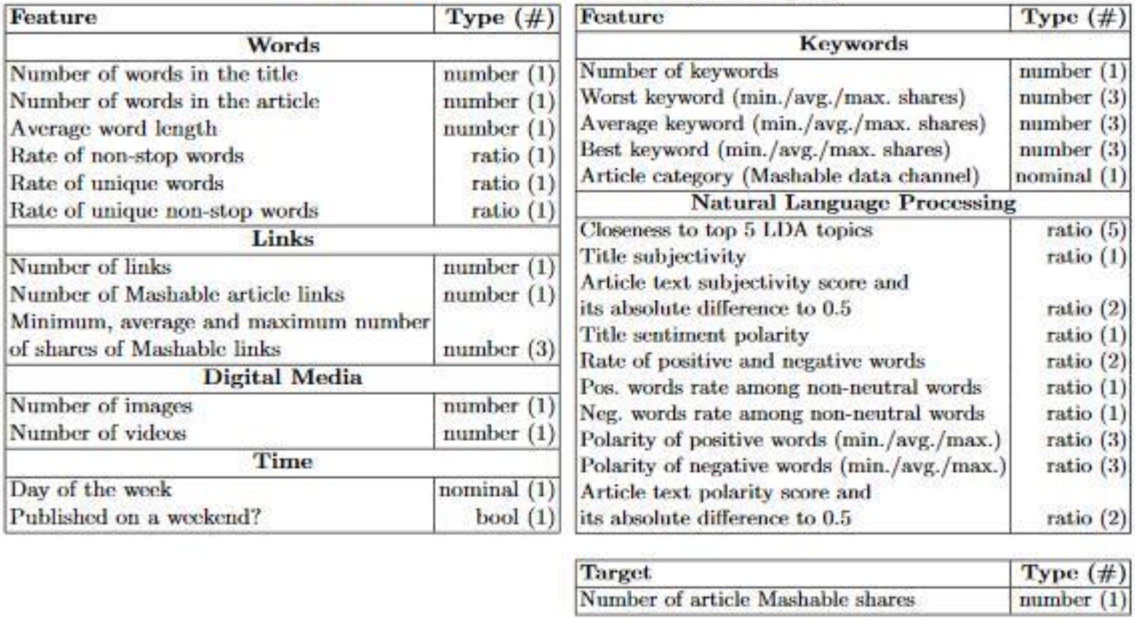
Table 1: Statistical measures of the Mashable dataset.



		Articles per day			
Number of articles	**Total days**	**Average**	**Standard Deviation**	**Min**	**Max**
39,000	709	55.00	22.65	12	105

A large list of characteristics has selected that describe different aspects of the article and that are considered possibly relevant to influence the number of shares. Some of the features are dependent of particularities of the Mashable service: articles often reference other articles published in the same service; and articles have meta-data, such as keywords, data channel type and total number of shares (when considering Facebook, Twitter, Google+, LinkedIn, Stumble-Upon and Pinterest). Thus, the minimum, average and maximum number of shares (known before publication) are extracted in all Mashable links cited in the article. Similarly, all article keyword average shares (known before publication) are ranked, in order to get the worst, average and best keywords. For each of these keywords, the minimum, average and maximum number of shares are extracted. The data channel categories are: “lifestyle”, “bus”, “entertainment”, “socmed”, “tech”, “viral” and “world”.

Table 2: List of attributes by category.



Feature	Type (#)
Words	
Number of words in the title	number (1)
Number of words in the article	number (1)
Average word length	number (1)
Rate of non-stop words	ratio (1)
Rate of unique words	ratio (1)
Rate of unique non-stop words	ratio (1)
Links	
Number of links	number (1)
Number of Mashable article links	number (1)
Minimum, average and maximum number of shares of Mashable links	number (3)
Digital Media	
Number of images	number (1)
Number of videos	number (1)
Time	
Day of the week	nominal (1)
Published on a weekend?	bool (1)

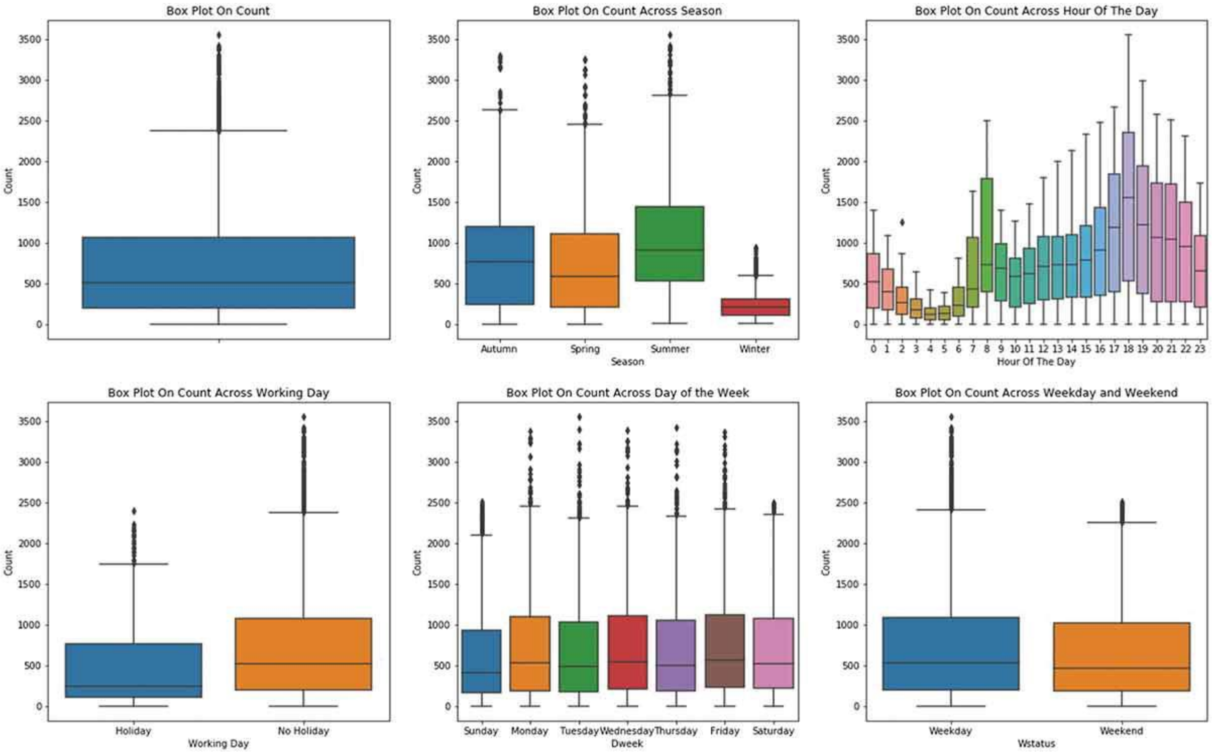
Feature	Type (#)
Keywords	
Number of keywords	number (1)
Worst keyword (min./avg./max. shares)	number (3)
Average keyword (min./avg./max. shares)	number (3)
Best keyword (min./avg./max. shares)	number (3)
Article category (Mashable data channel)	nominal (1)
Natural Language Processing	
Closeness to top 5 LDA topics	ratio (5)
Title subjectivity	ratio (1)
Article text subjectivity score and its absolute difference to 0.5	ratio (2)
Title sentiment polarity	ratio (1)
Rate of positive and negative words	ratio (2)
Pos. words rate among non-neutral words	ratio (1)
Neg. words rate among non-neutral words	ratio (1)
Polarity of positive words (min./avg./max.)	ratio (3)
Polarity of negative words (min./avg./max.)	ratio (3)
Article text polarity score and its absolute difference to 0.5	ratio (2)

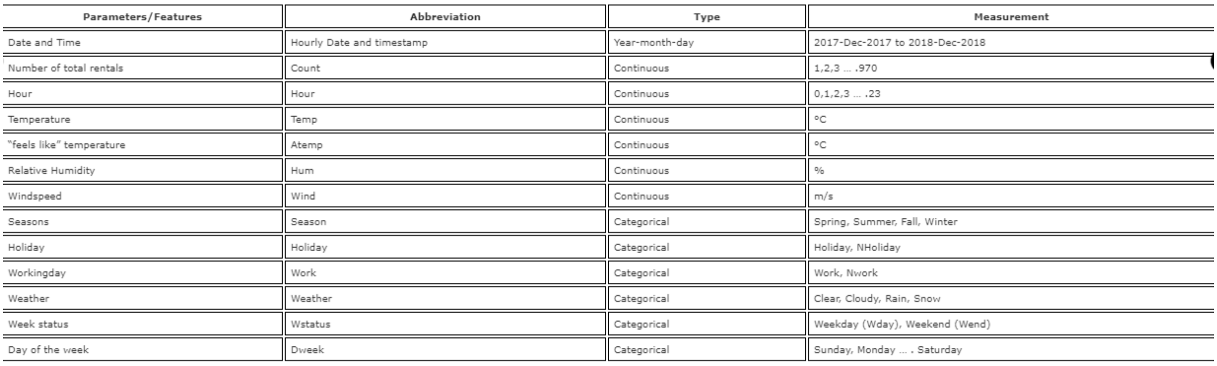
Target	Type (#)
Number of article Mashable shares	number (1)

Regression Dataset:

Seoul Bike Sharing Demand Data Set

This research paper presents a rule-based regression predictive model for bike sharing demand prediction. Both data have weather data associated with it for each hour. For both the dataset, five statistical models were trained with optimized hyperparameters using a repeated cross validation approach and testing set is used for evaluation.





Parameters/Features	Abbreviation	Type	Measurement
Date and Time	Hourly Date and timestamp	Year-month-day	2017-Dec-2017 to 2018-Dec-2018
Number of total rentals	Count	Continuous	1,2,3 … .970
Hour	Hour	Continuous	0,1,2,3 … .23
Temperature	Temp	Continuous	°C
"feels like" temperature	Atemp	Continuous	°C
Relative Humidity	Hum	Continuous	%
Windspeed	Wind	Continuous	m/s
Seasons	Season	Categorical	Spring, Summer, Fall, Winter
Holiday	Holiday	Categorical	Holiday, NHoliday
Workingday	Work	Categorical	Work, Nwork
Weather	Weather	Categorical	Clear, Cloudy, Rain, Snow
Week status	Wstatus	Categorical	Weekday (Wday), Weekend (Wend)
Day of the week	Dweek	Categorical	Sunday, Monday … . Saturday

The variable importance results have shown that Temperature and Hour of the day are the most influential variables in the hourly rental bike demand prediction.