Assignment Submission Report

Data Mining & Warehousing Assignment 1

IT13011130 – Gunathilaka D. D. T. M.(08/22/2016)

The following report describes the procedure that has been followed while implementing data mining model for Assignment 1 of Data Mining & Warehousing module.

Declaration

I hear by declare that following task is my own creation and it does not violate any constraints of the assignment.

Gunathilaka D. D. T. M. (IT13011130)

Table of Contents

Introduction

Case Study

In this assignment it has selected a data set which summarizes a heterogeneous set of features about articles published by Mashable in a period of two years. (www.mashable.com). The respective data set consisted, 39797 records with 60 columns. Each row is identified by the URL of the article and the dependent variable (shares) is continuously valued. The goal is to predict the number of shares in social networks (popularity). In other words, we need to find what kind of articles are attracting eyes from general population using the other 58 independent attributes. In the implemented solution it trains a model to predict the number of shares. Therefore "Regression" is used as the data mining technique.

Background Study for Regression

Regression is a "Predictive" data-mining technique that falls under "Supervised Learning "category. In Oracle there are 2 algorithms to implement regression based data mining models. Both algorithms are particularly suited for mining data sets that have very high dimensionality (many attributes), including transactional and unstructured data. In the model build (training) process, a regression algorithm estimates the value of the target as a function of the predictors for each case in the build data. These relationships between predictors and target are summarized in a model, which can then be applied to a different data set in which the target values are unknown.

Generalized Linear Models (GLM)

GLM is a popular statistical technique for linear modeling. Oracle Data Mining implements GLM for regression and for binary classification.

GLM provides extensive coefficient statistics and model statistics, as well as row diagnostics. GLM also supports confidence bounds.

Support Vector Machines (SVM)

SVM is a powerful, state-of-the-art algorithm for linear and nonlinear regression. Oracle Data Mining implements SVM for regression and other mining functions.

SVM regression supports two kernels: The Gaussian kernel for nonlinear regression, and the linear kernel for linear regression. SVM also supports active learning.

Furthermore, Root Mean Squared Error and the Mean Absolute Error are commonly used statistics for evaluating the overall quality of a regression model.

Root Mean Squared Error

The Root Mean Squared Error (RMSE) is the square root of the average squared distance of a data point from the fitted line.

This SQL expression calculates the RMSE.

*SQRT(AVG((predicted_value - actual_value) * (predicted_value - actual_value)))*

Mean Absolute Error

The Mean Absolute Error (MAE) is the average of the absolute value of the residuals (error). The MAE is very similar to the RMSE but is less sensitive to large errors.

This SQL expression calculates the MAE.

AVG(ABS(predicted_value - actual_value))

Procedure

Data Cleaning

Before starting the implementation, data preprocessed by adjusting data type and excluding URL as predictor a column. "URL" is also the only unique field in the data set which can be used as case id. Unnecessary data needs to be removed from the original data set, since they might cause to increase the error-rate of predictions. Therefore, rows with empty data were removed from original data set.

Data Integration

The original data set was in .CSV format. Once it has been cleansed as next step; "ONLINENEWSPOPULARITY" table was created under Oracle user account which has data mining privileges. Since some attribute data had around 12 decimal points "NUMBER (38,20)" used for their column type to preserve data.

Implementation

There are 2 ways to implement the models.

1. Using query
2. Using SQL developer GUI

1st approach

Settings for the selected miner models are stored in a table as key-value pairs. Only the column names have to be same. (setting name, setting value)

```
CREATE TABLE miner_model_settings(setting_name VARCHAR2(30),setting_value VARCHAR2(4000));
/
```

As next step for this example, the settings of the best regression algorithm for the dataset(GLM) with feature selection with 1.3345% predictive confidence were stored in above created table.

```
BEGIN
INSERT INTO MINER_MODEL_SETTINGS VALUES(
DBMS_DATA_MINING.ALGO_NAME,DBMS_DATA_MINING.ALGO_GENERALIZED_LINEAR_MODEL
);
INSERT INTO MINER_MODEL_SETTINGS VALUES(
DBMS_DATA_MINING.GLMS_RIDGE_REGRESSION,DBMS_DATA_MINING.GLMS_RIDGE_REG_DISABLE
);
INSERT INTO MINER_MODEL_SETTINGS VALUES(
DBMS_DATA_MINING.GLMS_FTR_SELECTION,DBMS_DATA_MINING.GLMS_FTR_SELECTION_ENABLE
);
INSERT INTO MINER_MODEL_SETTINGS VALUES(
DBMS_DATA_MINING.PREP_AUTO,DBMS_DATA_MINING.PREP_AUTO_ON
);
INSERT INTO MINER_MODEL_SETTINGS
VALUES(DBMS_DATA_MINING.GLMS_DIAGNOSTICS_TABLE_NAME,'ROW_DIAGNOSTIC_STATISTICS');
COMMIT;
END;
/
```

Following query is executed to build and train the model. In below query ‘case_id_column_name’ is a unique key which used to identify each record separately while ‘target_column_name’ is the target that is required to predict by the model.

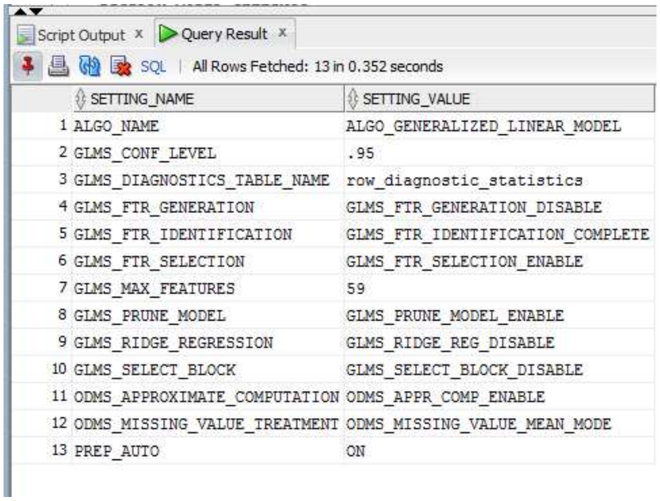
```
BEGIN
DBMS_DATA_MINING.CREATE_MODEL
(
MODEL_NAME => 'FINAL_GLM_MODEL',
MINING_FUNCTION => DBMS_DATA_MINING.REGRESSION,
DATA_TABLE_NAME     => 'ONLINENEWSPOPULARITY',
CASE_ID_COLUMN_NAME => 'URL',
TARGET_COLUMN_NAME  => 'SHARES',
SETTINGS_TABLE_NAME => 'MINER_MODEL_SETTINGS'
);
COMMIT;
END;
/
```

When PL/SQL procedure successfully completed, in order to test the created model “ONLINENEWSPOPULARITY” table is used. Therefore, following query is executed to test the implemented model.

```
SELECT URL,SHARES, PREDICTION(FINAL_GLM_MODEL USING *) PRED FROM ONLINENEWSPOPULARITY;
```

Model Settings:

```
COLUMN SETTING_NAME FORMAT A30
COLUMN SETTING_VALUE FORMAT A30
SELECT SETTING_NAME, SETTING_VALUE  FROM USER_MINING_MODEL_SETTINGS WHERE MODEL_NAME =
FINAL_GLM_MODEL' ORDER BY SETTING_NAME;
```

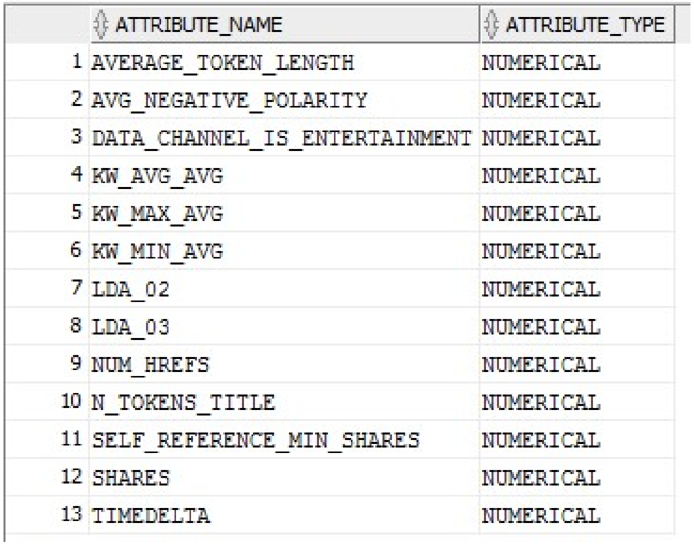


All Rows Fetched: 13 in 0.352 seconds

	SETTING_NAME	SETTING_VALUE
1	ALGO_NAME	ALGO_GENERALIZED_LINEAR_MODEL
2	GLMS_CONF_LEVEL	.95
3	GLMS_DIAGNOSTICS_TABLE_NAME	row_diagnostic_statistics
4	GLMS_FTR_GENERATION	GLMS_FTR_GENERATION_DISABLE
5	GLMS_FTR_IDENTIFICATION	GLMS_FTR_IDENTIFICATION_COMPLETE
6	GLMS_FTR_SELECTION	GLMS_FTR_SELECTION_ENABLE
7	GLMS_MAX_FEATURES	59
8	GLMS_PRUNE_MODEL	GLMS_PRUNE_MODEL_ENABLE
9	GLMS_RIDGE_REGRESSION	GLMS_RIDGE_REG_DISABLE
10	GLMS_SELECT_BLOCK	GLMS_SELECT_BLOCK_DISABLE
11	ODMS_APPROXIMATE_COMPUTATION	ODMS_APPR_COMP_ENABLE
12	ODMS_MISSING_VALUE_TREATMENT	ODMS_MISSING_VALUE_MEAN_MODE
13	PREP_AUTO	ON

Model Signature(Attributes):

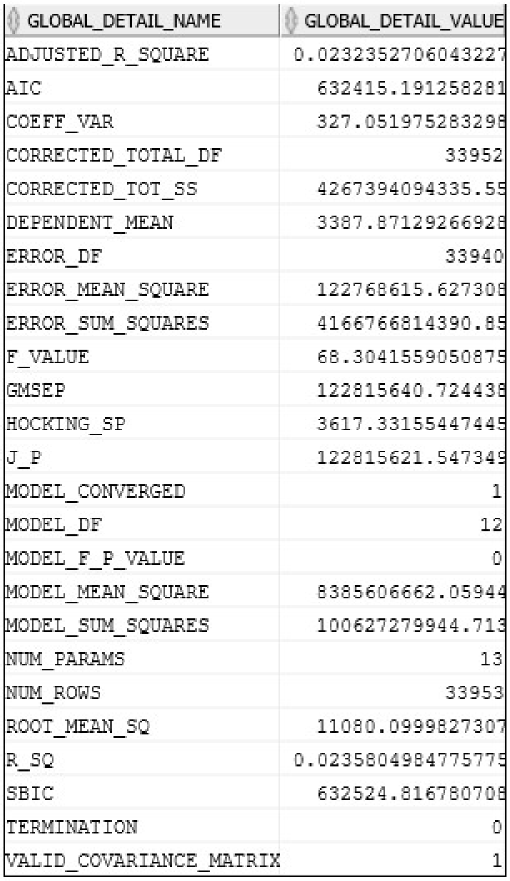
```
COLUMN ATTRIBUTE_NAME FORMAT A40
COLUMN ATTRIBUTE_TYPE FORMAT A20
SELECT ATTRIBUTE_NAME, ATTRIBUTE_TYPE  FROM USER_MINING_MODEL_ATTRIBUTES WHERE MODEL_NAME =
'FINAL_GLM_MODEL'ORDER BY ATTRIBUTE_NAME;
```



	ATTRIBUTE_NAME	ATTRIBUTE_TYPE
1	AVERAGE_TOKEN_LENGTH	NUMERICAL
2	AVG_NEGATIVE_POLARITY	NUMERICAL
3	DATA_CHANNEL_IS_ENTERTAINMENT	NUMERICAL
4	KW_AVG_AVG	NUMERICAL
5	KW_MAX_AVG	NUMERICAL
6	KW_MIN_AVG	NUMERICAL
7	LDA_02	NUMERICAL
8	LDA_03	NUMERICAL
9	NUM_HREFS	NUMERICAL
10	N_TOKENS_TITLE	NUMERICAL
11	SELF_REFERENCE_MIN_SHARES	NUMERICAL
12	SHARES	NUMERICAL
13	TIMEDELTA	NUMERICAL

Global statistics:

```
SELECT *  FROM TABLE(DBMS_DATA_MINING.GET_MODEL_DETAILS_GLOBAL('FINAL_GLM_MODEL'))ORDER BY
GLOBAL_DETAIL_NAME;
```



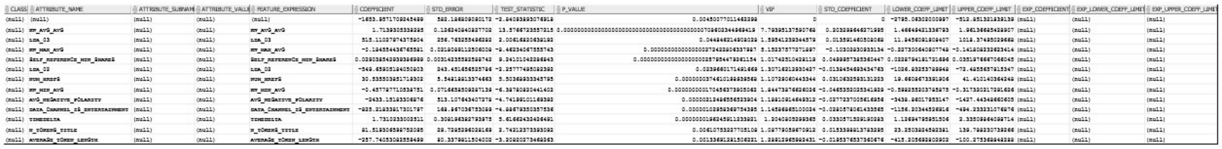
GLOBAL_DETAIL_NAME	GLOBAL_DETAIL_VALUE
ADJUSTED_R_SQUARE	0.0232352706043227
AIC	632415.191258281
COEFF_VAR	327.051975283298
CORRECTED_TOTAL_DF	33952
CORRECTED_TOT_SS	4267394094335.55
DEPENDENT_MEAN	3387.87129266928
ERROR_DF	33940
ERROR_MEAN_SQUARE	122768615.627308
ERROR_SUM_SQUARES	4166766814390.85
F_VALUE	68.3041559050875
GMSEP	122815640.724438
HOCKING_SP	3617.33155447445
J_P	122815621.547349
MODEL_CONVERGED	1
MODEL_DF	12
MODEL_F_P_VALUE	0
MODEL_MEAN_SQUARE	8385606662.05944
MODEL_SUM_SQUARES	100627279944.713
NUM_PARAMS	13
NUM_ROWS	33953
ROOT_MEAN_SQ	11080.0999827307
R_SQ	0.0235804984775775
SBIC	632524.816780708
TERMINATION	0
VALID_COVARIANCE_MATRIX	1

GLM statistics | Coefficient statistics | Features and their p_values:

```
SELECT * FROM TABLE (DBMS_DATA_MINING.GET_MODEL_DETAILS_GLM('FINAL_GLM_MODEL'));
SELECT * FROM TABLE(DBMS_DATA_MINING.GET_MODEL_DETAILS_GLM(MODEL_NAME =>
'FINAL_GLM_MODEL'));

SET LINE 120
COLUMN FEATURE_EXPRESSION FORMAT A53
SELECT FEATURE_EXPRESSION, COEFFICIENT, STD_ERROR, TEST_STATISTIC,
  P_VALUE, STD_COEFFICIENT, LOWER_COEFF_LIMIT, UPPER_COEFF_LIMIT
  FROM TABLE(DBMS_DATA_MINING.GET_MODEL_DETAILS_GLM('FINAL_GLM_MODEL'));


SET LIN 80
SET PAGES 20
SELECT FEATURE_EXPRESSION, COEFFICIENT, P_VALUE
  FROM TABLE(DBMS_DATA_MINING.GET_MODEL_DETAILS_GLM('FINAL_GLM_MODEL'))
  ORDER BY P_VALUE;
```



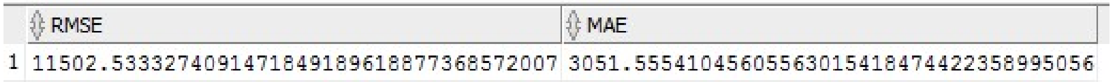
CLASS	ATTRIBUTE_NAME	ATTRIBUTE_SUBNAME	ATTRIBUTE_VALUE	FEATURE_EXPRESSION	COEFFICIENT	STD_ERROR	TEST_STATISTIC	P_VALUE	VIF	STD_COEFFICIENT	LOWER_COEFF_LIMIT	UPPER_COEFF_LIMIT	EXP_COEFFICIENT	EXP_LOWER_COEFF_LIMIT	EXP_UPPER_COEFF_LIMIT
(null)	(null)	(null)	(null)	(null)	-1653.9571709245499	582.18690090172	-2.84093893076918	0.00450077011462398	0	0	-2795.06302000997	-512.851321839139	(null)	(null)	(null)
(null)	KW_AVG_AVG	(null)	(null)	KW_AVG_AVG	1.7139305338285	0.126240940827702	13.5766723557215	0.000000000000000000000000000000000000000708602344969419	7.79395137590769	0.203289646271995	1.46649421336793	1.96136695428907	(null)	(null)	(null)
(null)	LDA_03	(null)	(null)	LDA_03	515.11027974375904	256.763255496282	2.00616820638192	0.044846214908029	1.59541239344579	0.013591460509062	11.8456091908407	1018.37495029669	(null)	(null)	(null)
(null)	KW_MAX_AVG	(null)	(null)	KW_MAX_AVG	-0.184554436765581	0.0219099112506009	-8.46234067555743	0.0000000000000000272422806337987	5.15237577271897	-0.103029300933134	-0.227300640907749	-0.141808233623414	(null)	(null)	(null)
(null)	SELF_REFERENCE_MIN_SHARES	(null)	(null)	SELF_REFERENCE_MIN_SHARES	0.029038542939396999	0.00314235582598742	9.24101042286843	0.00000000000000000257954479261154	1.01742510428119	0.0499957385360447	0.0228794191721696	0.0351976667066045	(null)	(null)	(null)
(null)	LDA_02	(null)	(null)	LDA_02	-549.65905184050903	243.451656525796	-2.25777495082392	0.0239660171491669	1.30716212930427	-0.0139454693434763	-1026.83253789948	-72.4855657815347	(null)	(null)	(null)
(null)	NUM_HREFS	(null)	(null)	NUM_HREFS	30.535503951719302	5.54818913374663	5.50369933345795	0.0000000374610198939569	1.10729060443344	0.0310632593131223	19.6608673391906	41.410140364249	(null)	(null)	(null)
(null)	KW_MIN_AVG	(null)	(null)	KW_MIN_AVG	-0.457787710538751	0.0716658509227139	-6.38780820441402	0.000000000170456373905062	1.94473976626026	-0.0465352025341939	-0.598255203785975	-0.317320217391626	(null)	(null)	(null)
(null)	AVG_NEGATIVE_POLARITY	(null)	(null)	AVG_NEGATIVE_POLARITY	-2433.15193306976	513.107643407879	-4.74199101169392	0.00000212496556523904	1.19910914649312	-0.0277237005616956	-3439.86017953147	-1427.44349660605	(null)	(null)	(null)
(null)	DATA_CHANNEL_IS_ENTERTAINMENT	(null)	(null)	DATA_CHANNEL_IS_ENTERTAINMENT	-925.21833917301797	168.967036753089	-4.98679350357536	0.00000102959369754395	1.14596965100024	-0.0290579061433565	-1156.20344526916	-494.233231076976	(null)	(null)	(null)
(null)	TIMEDELTA	(null)	(null)	TIMEDELTA	1.7310233002511	0.308196382793975	5.61662432426491	0.0000000196245911233921	1.2040905399365	0.0330571539190283	1.12694795951506	2.33509964098714	(null)	(null)	(null)
(null)	N_TOKENS_TITLE	(null)	(null)	N_TOKENS_TITLE	81.519306598752095	29.7285296028169	2.74212373393092	0.00610753227705109	1.08779059670912	0.0153398913793295	23.2503924580391	139.788230739266	(null)	(null)	(null)
(null)	AVERAGE_TOKEN_LENGTH	(null)	(null)	AVERAGE_TOKEN_LENGTH	-257.74053092559499	80.3379911504002	-3.20820273469263	0.00133691281506221	1.29913965992431	-0.0195376537360676	-415.205692903902	-100.275369944289	(null)	(null)	(null)

Validation:

Root Mean Square Error - Sqrt(Mean((x - x')^2))

Mean Absolute Error - Mean(| (x - x') |)

```
 COLUMN RMSE FORMAT 9999.99
COLUMN MAE FORMAT 9999.99
SELECT SQRT(AVG((A.PRED - B.SHARES) * (A.PRED - B.SHARES))) RMSE,
      AVG(ABS(A.PRED - B.SHARES)) MAE
  FROM (SELECT URL,SHARES, PREDICTION(FINAL_GLM_MODEL USING *) PRED
        FROM ONLINENEWSPOPULARITY) A,
      ONLINENEWSPOPULARITY B
  WHERE A.URL = B.URL;
```



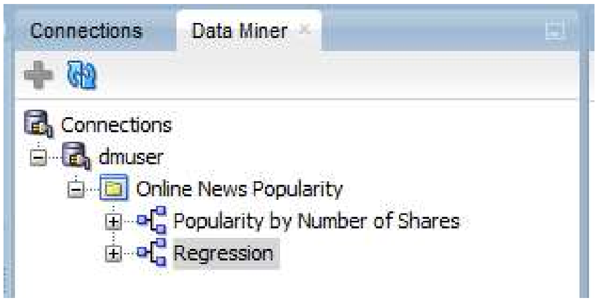
	RMSE	MAE
1	11502.53332740914718491896188773685 72007	3051.555410456055630154184744223589950 56

2nd approach

In the second method of model implementation a graphical user interface provided by Oracle SQL-Developer tool is used. Once the tool is opened following steps are followed to implement Classification mining model.

Steps

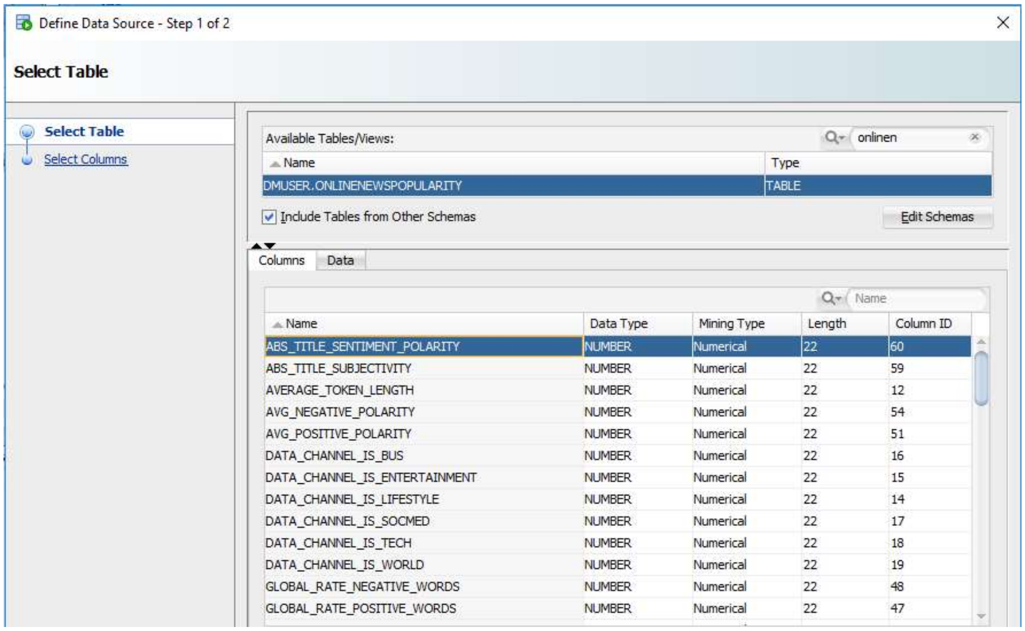
Create new project and work flow in created data miner user account under Oracle Data Miner. Name the project as "Online News Popularity" (any name) and define the workflow as "Popularity by Number of Shares" (any name)



Connections
- dmuser
  - Online News Popularity
    - Popularity by Number of Shares
    - Regression

Figure 1 Create Project & workflow

Next drag and drop Data Source from the tool palette to the workflow which is under "Data" section of palette. Once the "Data Source" node is added to workflow it will open a window to select a schema. This schema would be the table which provides training data to the model. Therefore, in our case select "DMUSER. ONLINENEWSPOPULARITY" as the schema. Since we have cleaned irrelevant columns previously all the existing columns of the table are required to build the model. Therefore, click "Finish"



Available Tables/Views: onlinen

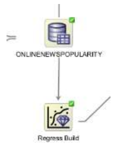
Name	Type
DMUSER.ONLINENEWSPOPULARITY	TABLE

☑ Include Tables from Other Schemas

Name	Data Type	Mining Type	Length	Column ID
ABS_TITLE_SENTIMENT_POLARITY	NUMBER	Numerical	22	60
ABS_TITLE_SUBJECTIVITY	NUMBER	Numerical	22	59
AVERAGE_TOKEN_LENGTH	NUMBER	Numerical	22	12
AVG_NEGATIVE_POLARITY	NUMBER	Numerical	22	54
AVG_POSITIVE_POLARITY	NUMBER	Numerical	22	51
DATA_CHANNEL_IS_BUS	NUMBER	Numerical	22	16
DATA_CHANNEL_IS_ENTERTAINMENT	NUMBER	Numerical	22	15
DATA_CHANNEL_IS_LIFESTYLE	NUMBER	Numerical	22	14
DATA_CHANNEL_IS_SOCMED	NUMBER	Numerical	22	17
DATA_CHANNEL_IS_TECH	NUMBER	Numerical	22	18
DATA_CHANNEL_IS_WORLD	NUMBER	Numerical	22	19
GLOBAL_RATE_NEGATIVE_WORDS	NUMBER	Numerical	22	48
GLOBAL_RATE_POSITIVE_WORDS	NUMBER	Numerical	22	47

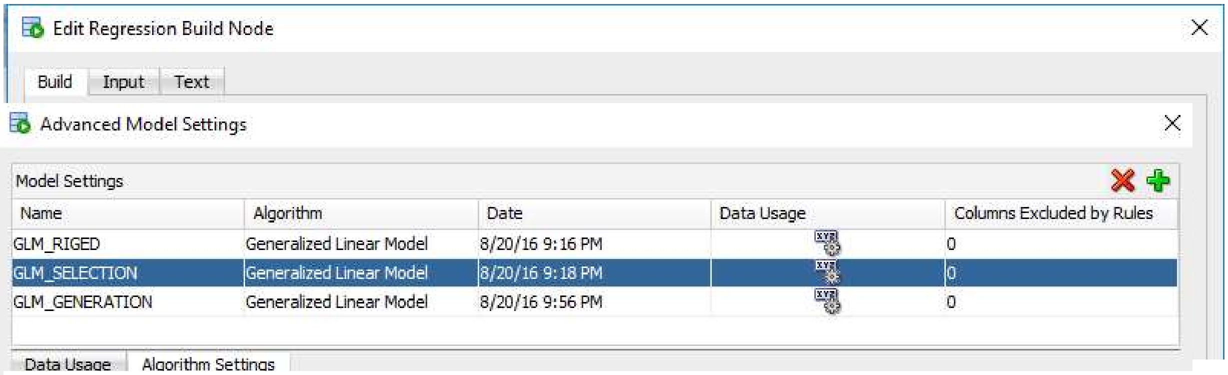
Figure 2 Select Data Source

As the next step the model node should be created. Therefore drag and drop Regression node from tool palette which is under "Models" section. Then link the two nodes (



)by using Link which is under "Linking Nodes" category or right click → connect.

Next double-click or right click → Edit and open edit window of "Regress Build" (i.e. Regression) node. Then use following setting:



Edit Regression Build Node

Build | Input | Text

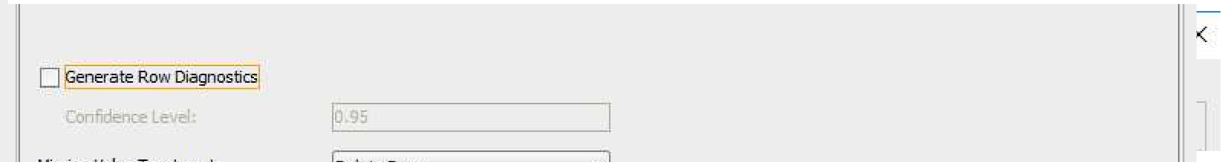
Advanced Model Settings

Model Settings

Name	Algorithm	Date	Data Usage	Columns Excluded by Rules
GLM_RIGED	Generalized Linear Model	8/20/16 9:16 PM		0
GLM_SELECTION	Generalized Linear Model	8/20/16 9:18 PM		0
GLM_GENERATION	Generalized Linear Model	8/20/16 9:56 PM		0

Data Usage | Algorithm Settings

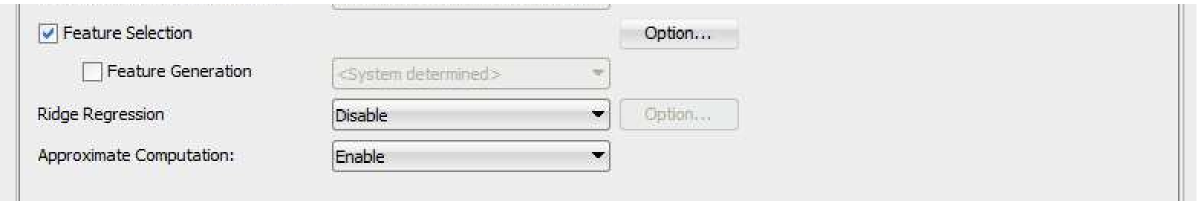
Figure 3 Add, Delete, Set Target and Case ID



□ Generate Row Diagnostics

Confidence Level: 0.95

Figure 4 Oracle will evaluate all attributes to determine which are important



☑ Feature Selection Option...

□ Feature Generation <System determined>

Ridge Regression Disable Option...

Approximate Computation: Enable

Figure 5 Advance settings for best confidence in dataset

Double Click on the model for advanced settings:

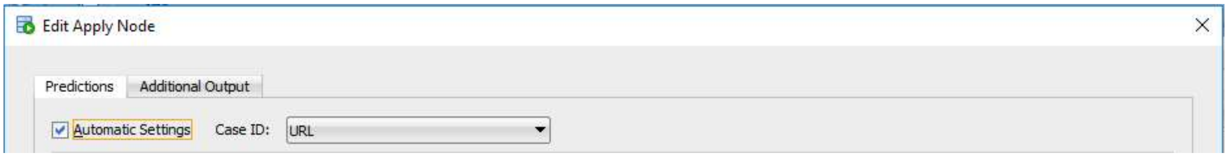
Press Ok twice to go back to work flow.

Now right click on regress build and click on Run button.

Once the model has successfully executed drag and drop "Apply" node and new "Data-Source" node from tool palette.

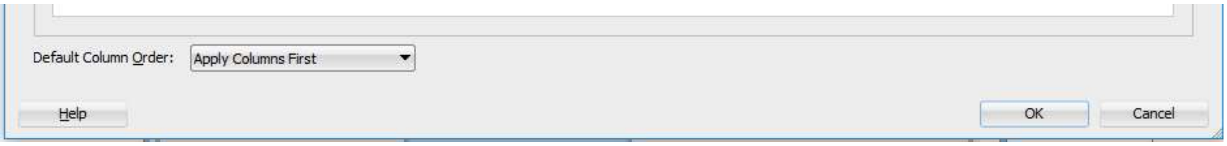
Next double click on newly added “Data-Source” node and “ONLINENEWSPOPULARITY” table in data source node.

In apply node:



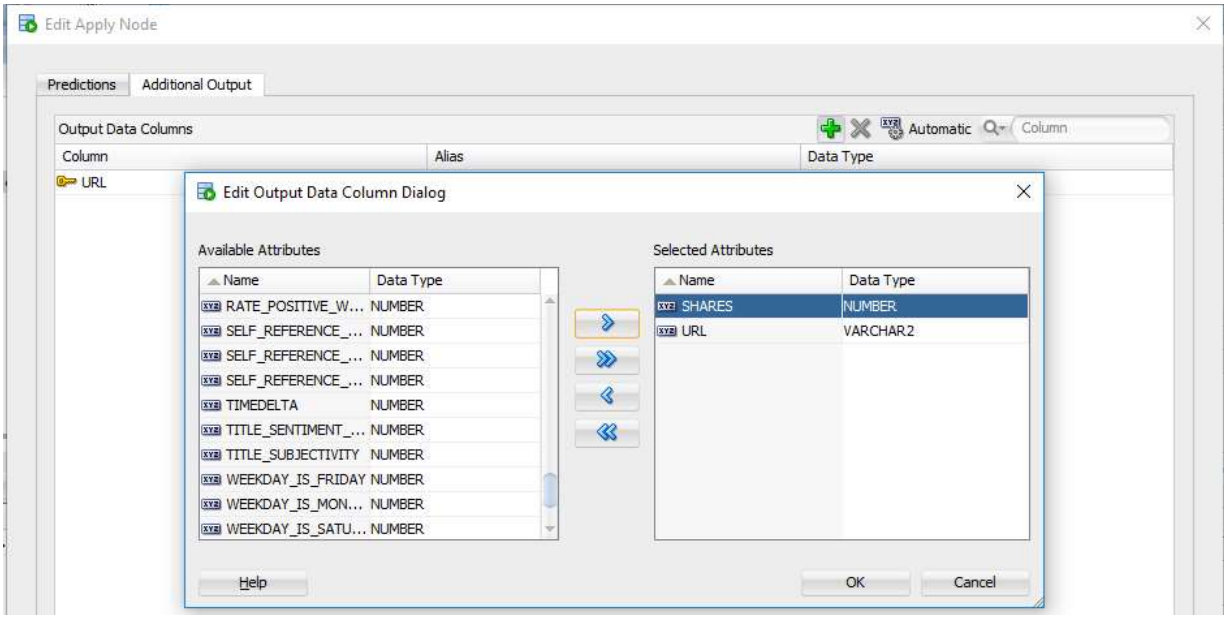
☑ Automatic Settings Case ID: URL

Figure 6 Set case ID for quicker predictions



Default Column Order: Apply Columns First

Figure 7 Set Column order



Column	Alias	Data Type
URL		

Available Attributes

Name	Data Type
RATE_POSITIVE_W...	NUMBER
SELF_REFERENCE_...	NUMBER
SELF_REFERENCE_...	NUMBER
SELF_REFERENCE_...	NUMBER
TIMEDELTA	NUMBER
TITLE_SENTIMENT_...	NUMBER
TITLE_SUBJECTIVITY	NUMBER
WEEKDAY_IS_FRIDAY	NUMBER
WEEKDAY_IS_MON...	NUMBER
WEEKDAY_IS_SATU...	NUMBER

Selected Attributes

Name	Data Type
SHARES	NUMBER
URL	VARCHAR2

Figure 8 Add Case Id as reference, actual shares to compare with predictions

Next re-run the entire workflow. This will apply implemented all algorithms to the records of “ONLINENEWSPOPULARITY” table. Now models are ready to provide predictions.

In order to display predictions a separate table called “REGRESSION_RESULT” is created. This table keeps the predictions, probability of those predictions and relevant case Ids made by models.

The implemented work-flow appears as below.

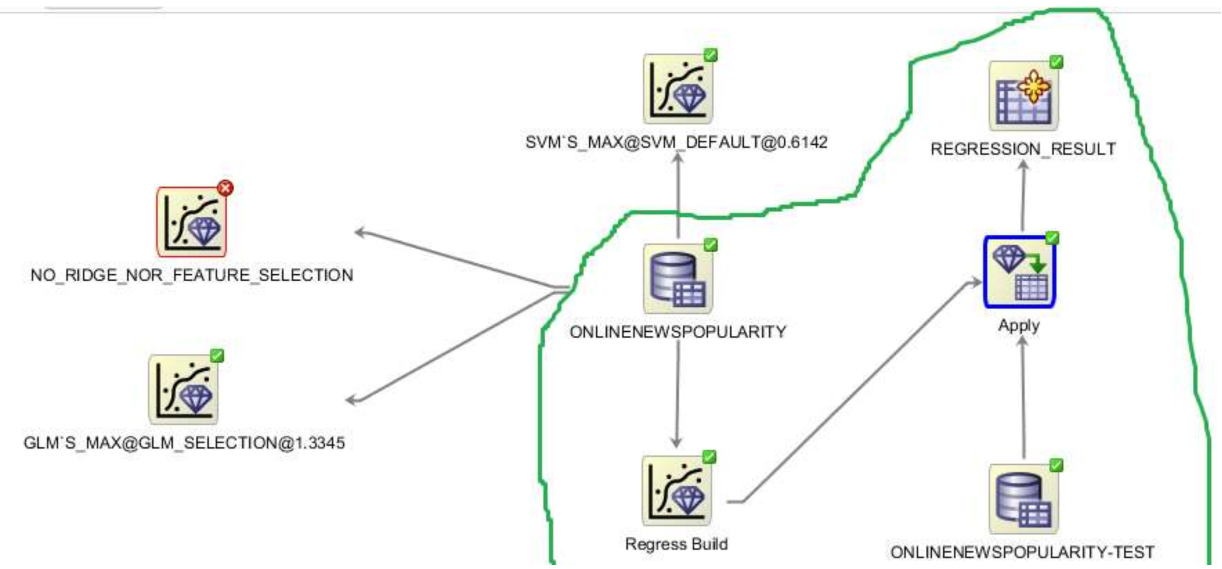

SVM`S_MAX@SVM_DEFAULT@0.6142
REGRESSION_RESULT
NO_RIDGE_NOR_FEATURE_SELECTION
ONLINENEWSPOPULARITY
Apply
GLM`S_MAX@GLM_SELECTION@1.3345
Regress Build
ONLINENEWSPOPULARITY-TEST

Figure 9 Selected Model for deploy

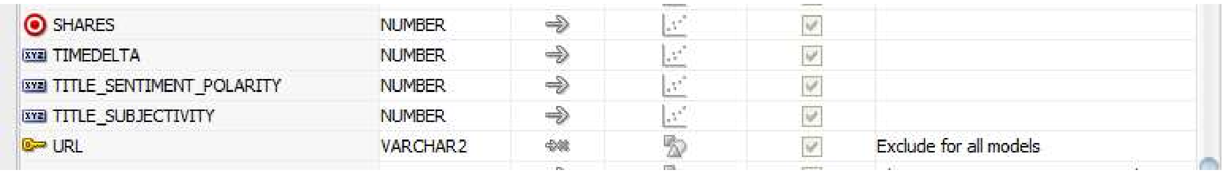
The accuracy of the models can be visualized by selecting "Compare Test Results" option in menu which appears once right click on "Class Build" node.

Right click on apply node → deploy → select node, dependent node and children node to generate query for the model.

In model settings:

Missing value treatment set to "Delete Row" to remove rows with null value

Shares (Independent variable) and URL not used in prediction

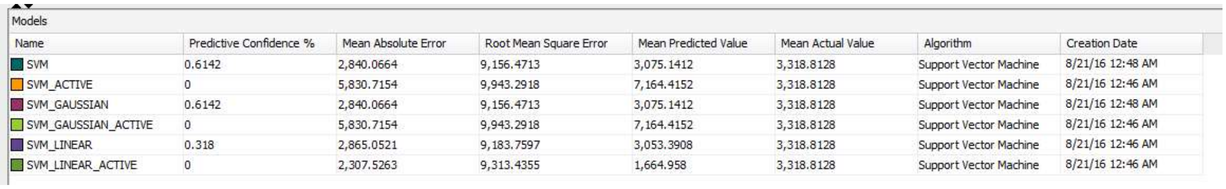


SHARES	NUMBER	☑	
TIMEDELTA	NUMBER	☑	
TITLE_SENTIMENT_POLARITY	NUMBER	☑	
TITLE_SUBJECTIVITY	NUMBER	☑	
URL	VARCHAR2	☑	Exclude for all models

Figure 10 URL and Shares attributes are not used

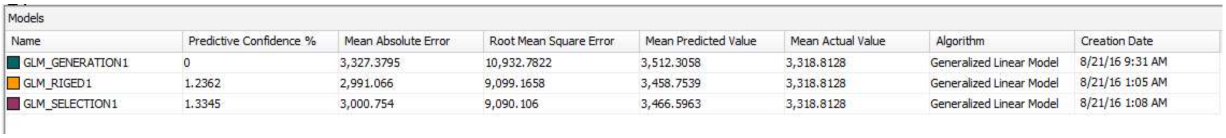
Results & Discussion

The overall performances of implemented models are shown below.



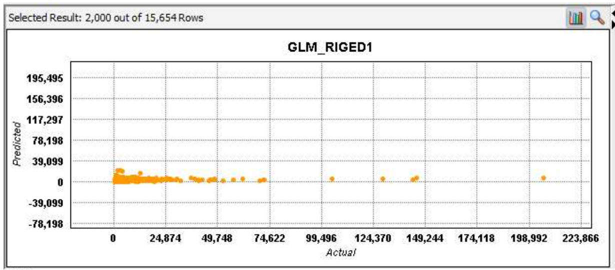
Models

Name	Predictive Confidence %	Mean Absolute Error	Root Mean Square Error	Mean Predicted Value	Mean Actual Value	Algorithm	Creation Date
■ SVM	0.6142	2,840.0664	9,156.4713	3,075.1412	3,318.8128	Support Vector Machine	8/21/16 12:48 AM
■ SVM_ACTIVE	0	5,830.7154	9,943.2918	7,164.4152	3,318.8128	Support Vector Machine	8/21/16 12:46 AM
■ SVM_GAUSSIAN	0.6142	2,840.0664	9,156.4713	3,075.1412	3,318.8128	Support Vector Machine	8/21/16 12:48 AM
■ SVM_GAUSSIAN_ACTIVE	0	5,830.7154	9,943.2918	7,164.4152	3,318.8128	Support Vector Machine	8/21/16 12:46 AM
■ SVM_LINEAR	0.318	2,865.0521	9,183.7597	3,053.3908	3,318.8128	Support Vector Machine	8/21/16 12:46 AM
■ SVM_LINEAR_ACTIVE	0	2,307.5263	9,313.4355	1,664.958	3,318.8128	Support Vector Machine	8/21/16 12:46 AM

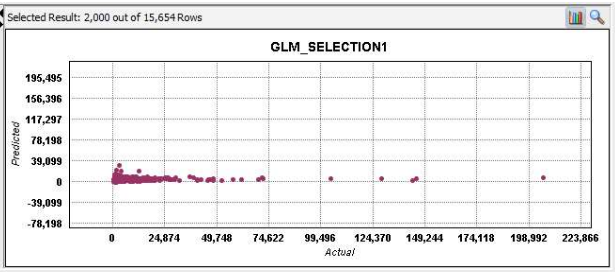


Models

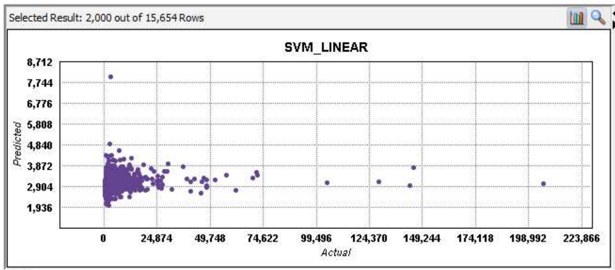
Name	Predictive Confidence %	Mean Absolute Error	Root Mean Square Error	Mean Predicted Value	Mean Actual Value	Algorithm	Creation Date
■ GLM_GENERATION1	0	3,327.3795	10,932.7822	3,512.3058	3,318.8128	Generalized Linear Model	8/21/16 9:31 AM
■ GLM_RIGED1	1.2362	2,991.066	9,099.1658	3,458.7539	3,318.8128	Generalized Linear Model	8/21/16 1:05 AM
■ GLM_SELECTION1	1.3345	3,000.754	9,090.106	3,466.5963	3,318.8128	Generalized Linear Model	8/21/16 1:08 AM



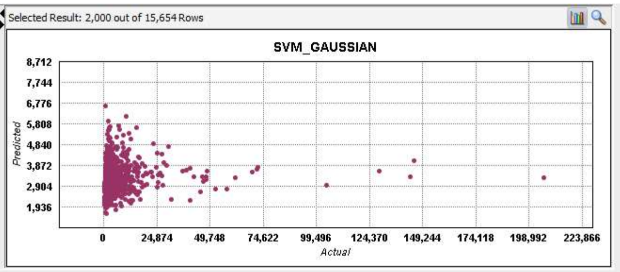
Selected Result: 2,000 out of 15,654 Rows



Selected Result: 2,000 out of 15,654 Rows

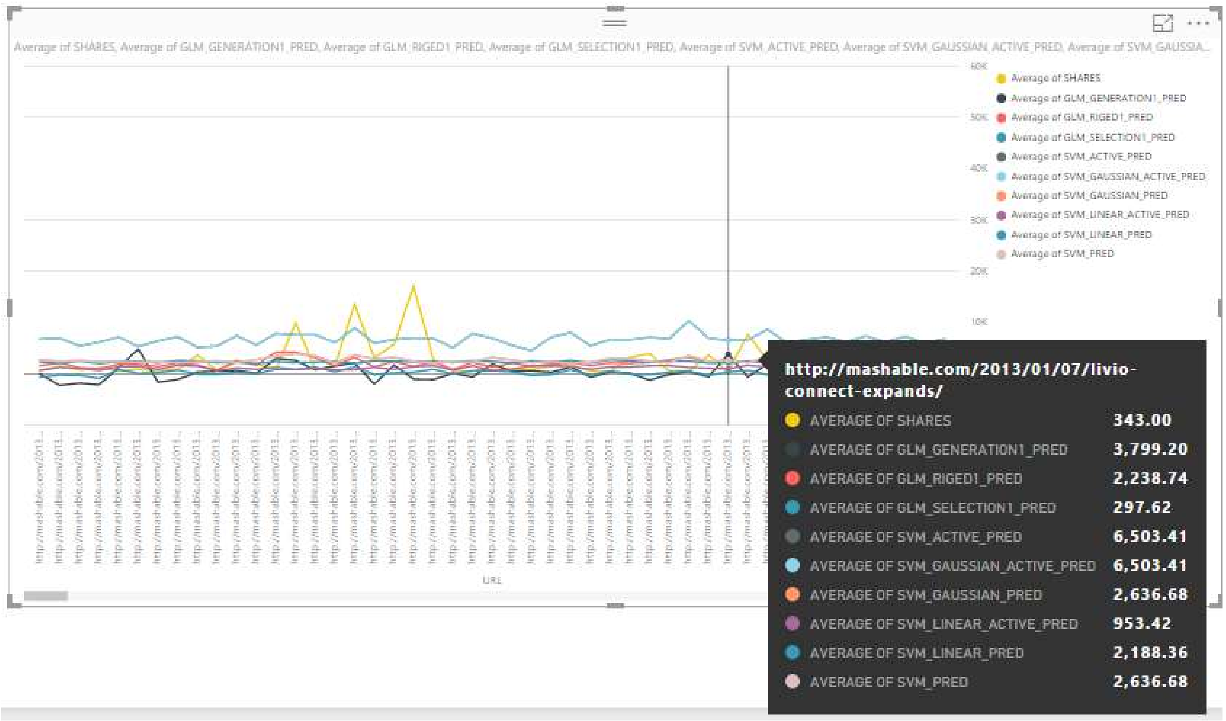


Selected Result: 2,000 out of 15,654 Rows



Selected Result: 2,000 out of 15,654 Rows

According to the above result SVM with linear kernel has the lowest spread and relatively highest accuracy. But when favoring the predictive confidence, we have to pick GLM with feature selection (confidence 0.318< 1.3345) as the best model.



http://mashable.com/2013/01/07/livio-connect-expands/

● AVERAGE OF SHARES	343.00
● AVERAGE OF GLM_GENERATION1_PRED	3,799.20
● AVERAGE OF GLM_RIGED1_PRED	2,238.74
● AVERAGE OF GLM_SELECTION1_PRED	297.62
● AVERAGE OF SVM_ACTIVE_PRED	6,503.41
● AVERAGE OF SVM_GAUSSIAN_ACTIVE_PRED	6,503.41
● AVERAGE OF SVM_GAUSSIAN_PRED	2,636.68
● AVERAGE OF SVM_LINEAR_ACTIVE_PRED	953.42
● AVERAGE OF SVM_LINEAR_PRED	2,188.36
● AVERAGE OF SVM_PRED	2,636.68

Figure 11 Prediction value distribution per post

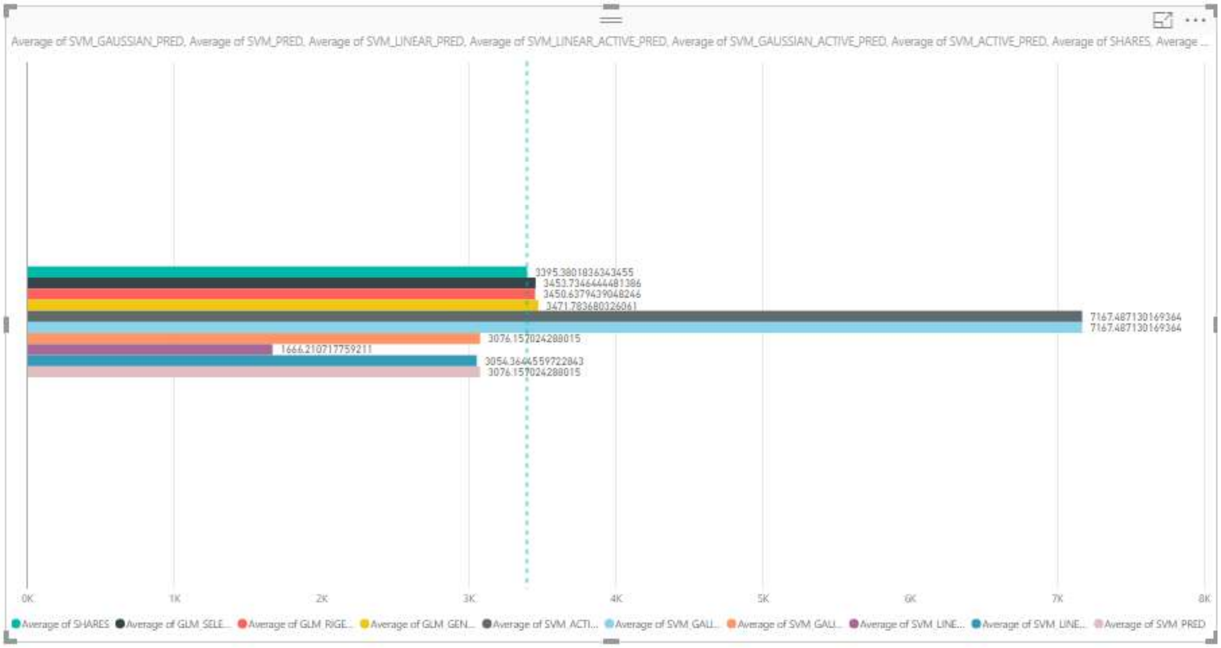
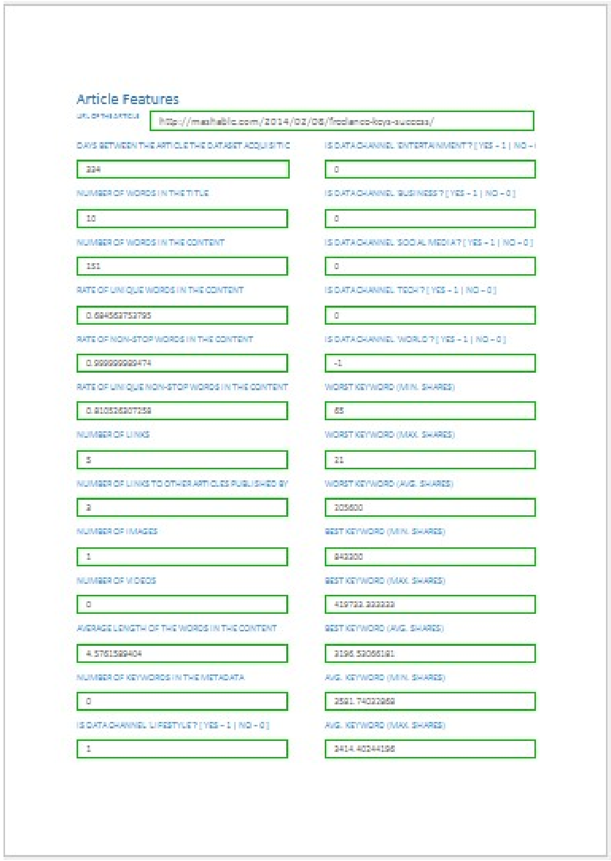


Figure 12 Overall Prediction value Average

Integration

The implemented models can be integrated with software components. Word form with VBA backend implemented for data insertion. Diagrams and visualization are displayed through an Excel instance but all the data manipulation done in the oracle server. Power BI used to visualize spread of training data.



Article Features

URL OF THE ARTICLE: http://mashable.com/2014/02/06/freelance-keys-success/

DAYS BETWEEN THE ARTICLE THE DATASET ACQUISITIC: 324

NUMBER OF WORDS IN THE TITLE: 10

NUMBER OF WORDS IN THE CONTENT: 151

RATE OF UNIQUE WORDS IN THE CONTENT: 0.684563753795

RATE OF NON-STOP WORDS IN THE CONTENT: 0.999999989474

RATE OF UNIQUE NON-STOP WORDS IN THE CONTENT: 0.810526307259

NUMBER OF LINKS: 5

NUMBER OF LINKS TO OTHER ARTICLES PUBLISHED BY: 3

NUMBER OF IMAGES: 1

NUMBER OF VIDEOS: 0

AVERAGE LENGTH OF THE WORDS IN THE CONTENT: 4.5761589404

NUMBER OF KEYWORDS IN THE METADATA: 0

IS DATA CHANNEL 'LIFESTYLE'? [YES – 1 | NO – 0]: 1

IS DATA CHANNEL 'ENTERTAINMENT'? [YES – 1 | NO – (: 0

IS DATA CHANNEL 'BUSINESS'? [YES – 1 | NO – 0]: 0

IS DATA CHANNEL 'SOCIAL MEDIA'? [YES – 1 | NO – 0]: 0

IS DATA CHANNEL 'TECH'? [YES – 1 | NO – 0]: 0

IS DATA CHANNEL 'WORLD'? [YES – 1 | NO – 0]: -1

WORST KEYWORD (MIN. SHARES): 65

WORST KEYWORD (MAX. SHARES): 21

WORST KEYWORD (AVG. SHARES): 205600

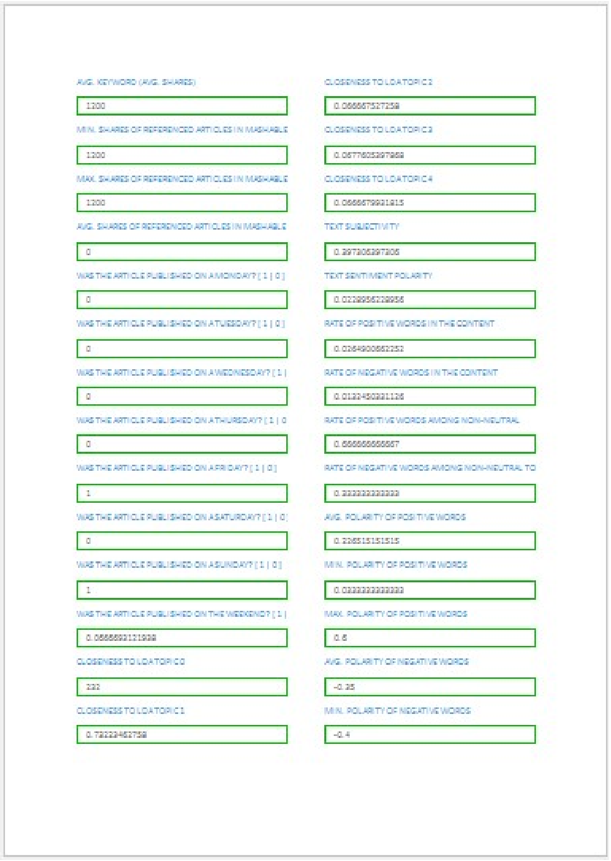
BEST KEYWORD (MIN. SHARES): 843300

BEST KEYWORD (MAX. SHARES): 419733.333333

BEST KEYWORD (AVG. SHARES): 3196.53066181

AVG. KEYWORD (MIN. SHARES): 2581.74022869

AVG. KEYWORD (MAX. SHARES): 2414.40244196



AVG. KEYWORD (AVG. SHARES): 1200

MIN. SHARES OF REFERENCED ARTICLES IN MASHABLE: 1200

MAX. SHARES OF REFERENCED ARTICLES IN MASHABLE: 1200

AVG. SHARES OF REFERENCED ARTICLES IN MASHABLE: 0

WAS THE ARTICLE PUBLISHED ON A MONDAY? [1 | 0]: 0

WAS THE ARTICLE PUBLISHED ON A TUESDAY? [1 | 0]: 0

WAS THE ARTICLE PUBLISHED ON A WEDNESDAY? [1 |: 0

WAS THE ARTICLE PUBLISHED ON A THURSDAY? [1 | 0: 0

WAS THE ARTICLE PUBLISHED ON A FRIDAY? [1 | 0]: 1

WAS THE ARTICLE PUBLISHED ON A SATURDAY? [1 | 0]: 0

WAS THE ARTICLE PUBLISHED ON A SUNDAY? [1 | 0]: 1

WAS THE ARTICLE PUBLISHED ON THE WEEKEND? [1 |: 0.0666692121938

CLOSENESS TO LDA TOPIC 0: 222

CLOSENESS TO LDA TOPIC 1: 0.72222462758

CLOSENESS TO LDA TOPIC 2: 0.066667527258

CLOSENESS TO LDA TOPIC 3: 0.0677605397969

CLOSENESS TO LDA TOPIC 4: 0.0666679931815

TEXT SUBJECTIVITY: 0.397306397306

TEXT SENTIMENT POLARITY: 0.0228956228956

RATE OF POSITIVE WORDS IN THE CONTENT: 0.0264900662252

RATE OF NEGATIVE WORDS IN THE CONTENT: 0.0132450331126

RATE OF POSITIVE WORDS AMONG NON-NEUTRAL: 0.666666666667

RATE OF NEGATIVE WORDS AMONG NON-NEUTRAL TO: 0.333333333333

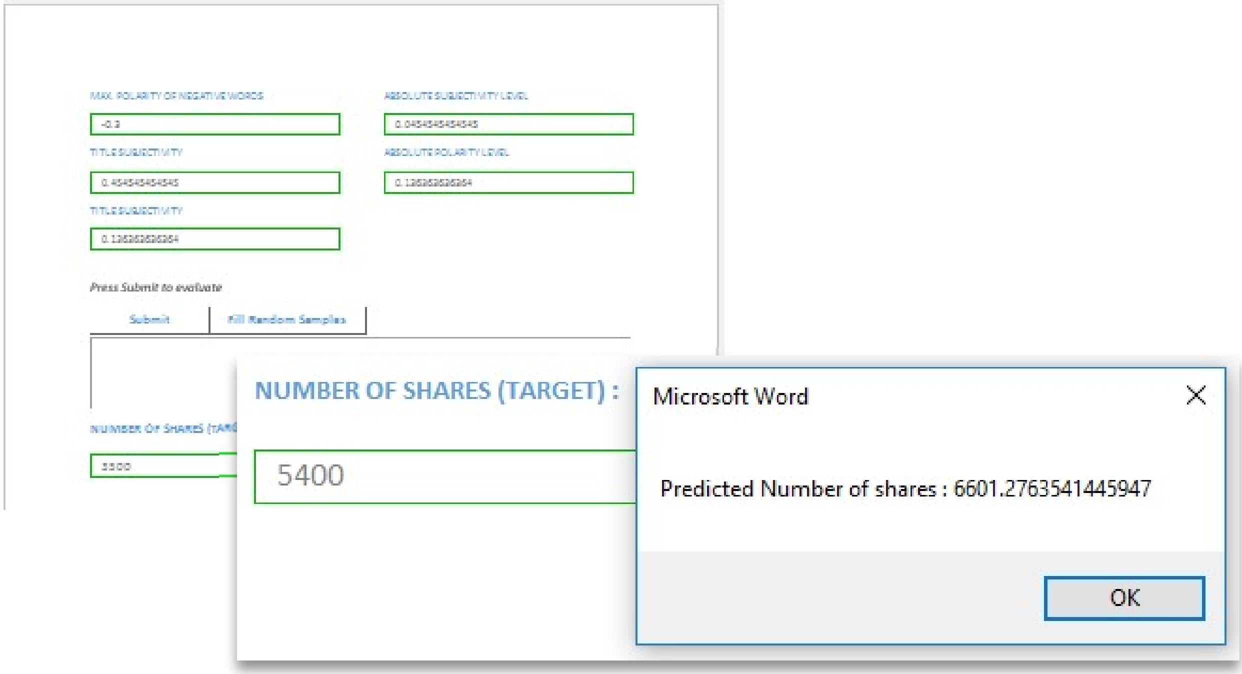
AVG. POLARITY OF POSITIVE WORDS: 0.226515151515

MIN. POLARITY OF POSITIVE WORDS: 0.0333333333333

MAX. POLARITY OF POSITIVE WORDS: 0.6

AVG. POLARITY OF NEGATIVE WORDS: -0.25

MIN. POLARITY OF NEGATIVE WORDS: -0.4



MAX. POLARITY OF NEGATIVE WORDS: -0.2

TITLE SUBJECTIVITY: 0.454545454545

TITLE SUBJECTIVITY: 0.136363636364

ABSOLUTE SUBJECTIVITY LEVEL: 0.0454545454545

ABSOLUTE POLARITY LEVEL: 0.136363636364

Press Submit to evaluate

NUMBER OF SHARES (TARG: 3300

NUMBER OF SHARES (TARGET) : 5400

Microsoft Word

Predicted Number of shares : 6601.2763541445947