This analysis report is presented to Best Melbourne Buys and summarizes the development, exploration and selection of predictive models for predicting price of the houses advertised for auction including the likely interest of the bidders that it will generate online.

Analysis engaged a sample dataset with both structured and unstructured data values of houses consisting 7206 properties sold between 2005 and 2015 is studied and refined for building predictive models.

Real-estate websites have been studied for the selection of considerable factors which determines the variation of price and page visits ("Property Data - Consumer Affairs Victoria") which identified as the target variables to predict. There are number of variable in the data set which does not have importance to consider due to vast distribution on data points with unique vales. Hence, those variables are not considered and reject to include as an input on model building. In addition, number of activates on cleaning the dataset has been executed which explains in data preparation section.

The text data, title and description of properties are considered for model building including other significant variables such as property type, bedrooms, bathrooms, car space and suburb to examine the importance of content which affects the price and interest of bidders.

During text analytics, using property description and title; it is constructed 10 set of words lists having 5 words per one set. Selections of these words are based on the frequency the word has appeared in the document and number of observations the word has appeared. In addition 7 group are constructed using these words and consider as input to understand the relationship with house price and the interest of the bidders upon page visits.

During model building activities, 3 different set of input data are considered and 3 different prediction models are built for comparison. This explains under the section “Predictive Models in EM”

Model building, using input variable from text analytics and other structured input variables provide the highest accuracy on predicting price over only structured input variables and only text input variables. The model predicts the price with the variation of $82000. As an example, if the predicted price of a property is $800000, the actual price is most likely to be in the range of $718000 to $882000. In addition, the model explained 80% of the variation in price, using both text input variables which derive from description and structured input variables. However, rest of the prediction models provide higher variation in predicting the price.

Model building using input variable from text analytics and other structured input variables provided the highest accuracy on predicting page visits over only structured input variables and only text input variables. The model predicts the likely interest of real-estate bidders with the variation of 674 visits. As an example, if the model predicts interest of 2000 visits, it can be stated that actually page visits will be more than 1328 visits.

As a recommendation, Best Melbourne Buys shall consider words for description, listed under “Generate Topics in Text Topic node” in page 6 to increase the price of the property. As an example, including words such as basement, stone, terrace, appliance, plan, complement, enhance, inviting appeal, cycle in the description of the property being putting for auction will help to increase the price. In addition, Best Melbourne Buys shall include words for title, listed under “Text analytic on Title variable” in page 7 to increase the price of the property. As an example, including words such sell, vendor, huge, heart, dream, bright, modern, delightful, spacious in the title of the property being putting for auction will help to increase the price.

Finally, the text inputs in this data set are more concentrated and found very effective during the whole analysis as well as final model building for better outcome. Specifically, property description is determined more significant and so analysed in depth. During text analytics process, subsequent dropping less significant words and combining words with more of similar meanings, has improved the model performance. Hence, there is possibility to increase the model accuracy by performing the above process iteratively.

However, it is recommended to revisit the model on latest text data for future predictions as bibber’s interest vary based on trend of real estate market.

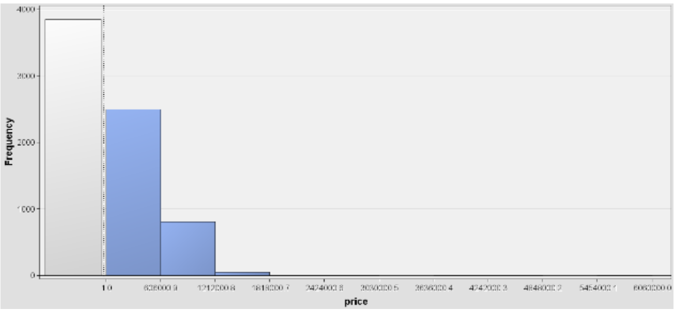
Data Preparation and Exploration in EM

At the very first glance towards the imported dataset in SAS EM, 7206 observations were identified. Observations which were having invalid attribute values; such as observations numbered 1446,1799,2100,3801 and 7015 were dropped as SAS did not allow data loading with those invalid text data in the variables.

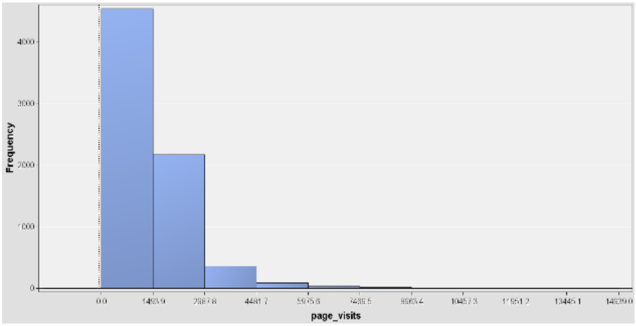
Secondly, the duplicate Property IDs such as 119732967,118388507 and 119732487 were dropped and the data set end up with 7198 observations.

There are almost 68% Properties are of Apartment types, 11.63% are Houses and 14.42% are Units. Thirdly, 6 properties with invalid property type such as ‘affords’, ‘Alfresco’, ‘fit’ and ‘you'll’ were dropped as it does not useful to include in the data set data set had 7192 observations.

Price and Page visits attributes were considered as target variables and have identified significant attributes for the analysis such as property type, bedrooms, bathrooms, car space and suburb. And according to the analysis, variables such as Agent Name, Agent Phone, Description, Latitude, Longitude, Property ID, Property Address, Sale Date and Title were dropped due to less significant for model building.



Looking towards visual representations, some useful insights were captured. Frequency plot of Price demonstrates that there are total 3687 observations with Missing price which is 51% of total observations. However, for some missing price properties, the sold price was stated in either description or title attribute. During this investigation, we could match the price for 67 properties which help to reduce the number of properties with missing values. , 808 properties are having price between $1212000 and $606000.



Frequency Plot of Page visits revealed that there are 63% properties having less than 1493 visits. Moreover interesting insight is only less than 1% Properties have more than 5975 visits. There are 0.08% properties do not have page visits data.

Row Labels	Count of property_type
Apartment	4971
Duplex/Semi-Detached	17
Flat	27
House	837
Other	7
Townhouse	196
Unit	1050
Villa	87
Grand Total	**7192**

figure DP_1

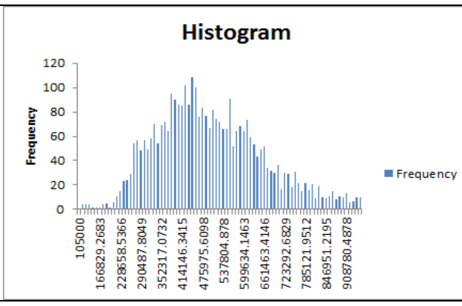


figure DP_2

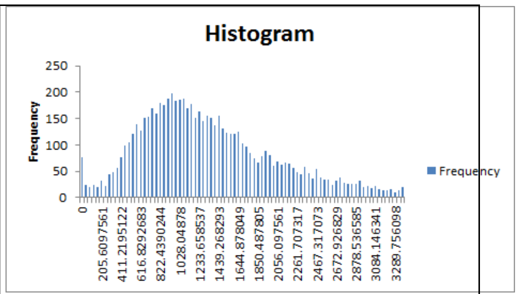
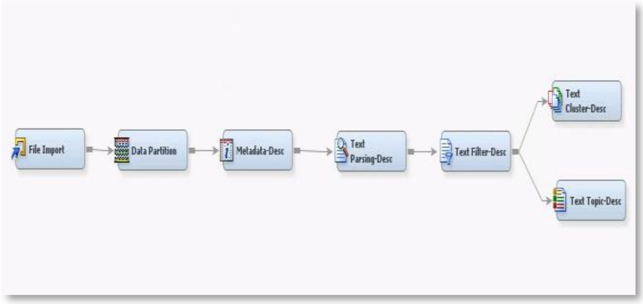


figure DP_3

The property type Retirement is appeared only once and there was a concern to club some other property types with major property types. Hence, property types, Retirement’, ‘Studio’, ‘Terrace’ and ‘UnitBlock’ has very less distribution and they were transformed to ‘Unit’, ‘Apartment’, ‘Townhouse’ and ‘unit’ respectively and the distribution of property type is as the figure DP_1.To have meaningful variability in sold price data, the outliers were identified through interquartile range. Based on the analysis 130 properties were dropped. Finally, the data set use to predict sold price consist of 7062 observations and the distribution of sold price is as the figure DP_2.Similarly, to have meaningful variability in page visit data, the outliers were identified through interquartile range. Based on the analysis 329 properties were dropped.Finally, the data set use to predict sold price consist of 6863 observations and the distribution of sold price is as the figure DP_3

Text Analytics in EM (Page 1)

Below section illustrates text analytic results using the data set on Melbourne house prices.



Using text analytics process flow, the unstructured text in both description and title attributes were converted to text topics and clusters. Based on these topics and clusters number of input variables were generated with data values which are significant to the target variables, price and page visits.
Modelling with text data by using text analytics to predict the house price gives significant weightage to the whole model and it helps to explains the variation in the target variable price.

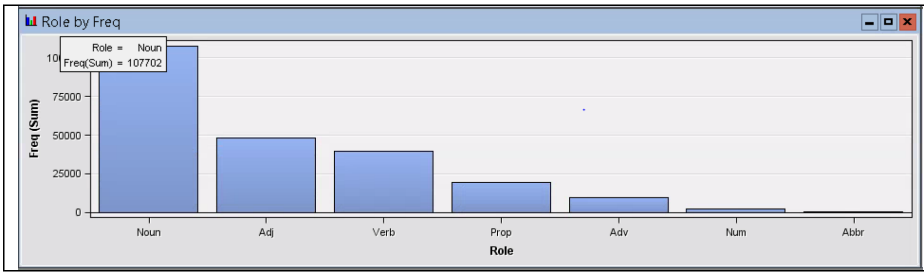
Working with "Terms" generated in Text Passing node for description variable

Terms were identified with frequency, a word occurs across all observations and the number of observation, which the word appears in the description.

Terms

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ kitchen	...Noun	Alpha	2797	2525	Y	+	1323	1
+ bedroom	...Noun	Alpha	2883	2290	Y	+	813	2
+ bathroom	...Noun	Alpha	2388	2179	Y	+	3615	3
+ apartment	...Noun	Alpha	2891	1974	Y	+	5261	4
+ laundry	...Noun	Alpha	1708	1648	Y	+	3534	5
living	...Noun	Alpha	1909	1572	Y		3094	6
+ dine	...Verb	Alpha	1663	1510	Y	+	2597	7
+ be	...Verb	Alpha	2610	1496	N	+	11296	8
+ floor	...Noun	Alpha	1662	1373	Y	+	1222	9
s	...Noun	Alpha	1948	1297	N		11115	10
+ room	...Noun	Alpha	1582	1213	Y	+	1729	11
bedroom	...Adj	Alpha	1277	1207	Y		5796	12

Words, such as "Apartment", "bedroom", "kitchen", "be" and "bathroom" were appearing in most of the description. The word "kitchen" is recorded 2797 times in 2525 observation and rank as first.



The description variable has recorded nouns 107702 time across all observations and very less abbreviations.

Managing "Terms" in Text Filter node to construct explainable topics for the model.

Based on text terms, respective frequency & weight for all the observations. Words occur less than 5 times across all observations were excluded.

Documents

Topic Weight	description	Observation Number
0.299	Perched high above the vibrant Ormond Road shops, restaurants and cafe's rests this stylish	199.0
0.294	This exceptionally larger than most 2 bedroom apartment is a must see. The light-filled	316.0
0.289	Modern, stylish and secure or great investment opportunity in most sought after location. Brand	664.0
0.28	Excellent room proportions are on offer in this oversized two bedroom apartment. Boasting	2639.0
0.278	Sleek and sophisticated with urban tones, this super spacious 2 bedroom 2 bathroom ground floor	2951.0
0.275	Evoking an exclusive sense of penthouse living, this prestigious, state-of-the-art apartment blends	362.0
0.274	Situated in the sought-after Elwood location, this spacious (approx 102 sqm) and sundrenched	1313.0
0.27	The beach and village at either end of the street! This near new, ultra-stylish, ground-floor,	1436.0
0.267	A truly spacious apartment in every sense of the word! With over 90sqm of internal living space	2625.0
0.267	Be the envy of all your friends with two private balconies ready for you to cook up those Summer	2261.0

The top descriptions, which against the generated 10 topics is belongs to the observation number 199.
The term with high weight based on numbers of times it occurs in all observations and number of observations it occurs might be occur in observation 199.

Text Analytics in EM (Page 2)

Terms

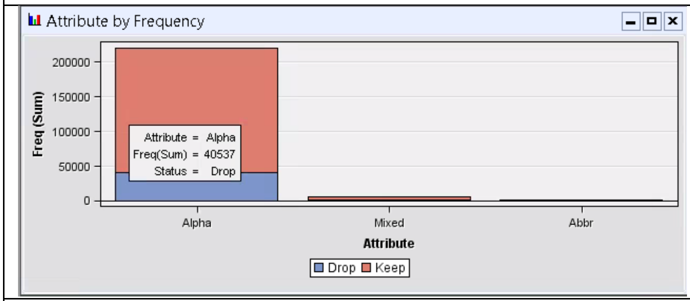
TERM	FREQ ▼	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
⊞ apartment	2891	1974	☐	0.0	Noun	Alpha
⊞ bedroom	2887	2293	☐	0.0	Noun	Alpha
⊞ kitchen	2798	2526	☐	0.0	Noun	Alpha
⊞ be	2615	1499	☐	0.0	Verb	Alpha
⊞ bathroom	2388	2179	☐	0.0	Noun	Alpha
⊞ kilda	2192	765	☑	0.184	Prop	Alpha
s	1948	1297	☐	0.0	Noun	Alpha
living	1909	1572	☑	0.082	Noun	Alpha
⊞ laundry	1708	1648	☑	0.069	Noun	Alpha
⊞ floor	1665	1374	☑	0.099	Noun	Alpha
⊞ dine	1663	1510	☑	0.083	Verb	Alpha
⊞ room	1589	1218	☑	0.117	Noun	Alpha
⊞ area	1506	1198	☑	0.118	Noun	Alpha
bedroom	1277	1207	☐	0.0	Adj	Alpha
separate	1255	987	☑	0.142	Adj	Alpha
⊞ park	1146	1037	☑	0.13	Verb	Alpha
spacious	1095	955	☑	0.142	Adj	Alpha

Word with high frequency implies that those words were having very less importance and the weights were set to 0. Hence those words were left from the term list.In addition,
words such as “birs” “herf”,”rel” and ”br” also left from the list as they did not have any meaning and invalid words which might affect to build top topics.

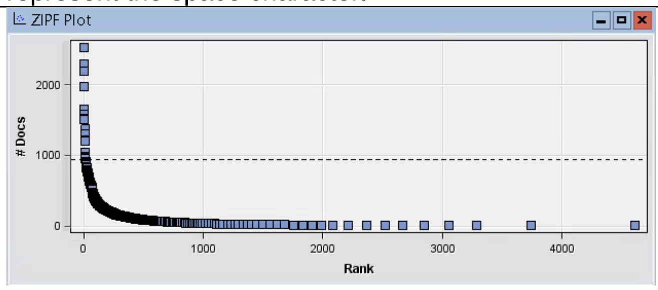
Terms

TERM	FREQ	# DOCS	KEEP	WEIGHT ▼	ROLE	ATTRIBUTE
nbsp	50	6	☐	0.893	Noun	Alpha
hodges	12	5	☐	0.815	Prop	Alpha
easterly	7	5	☐	0.814	Adj	Alpha
⊞ tour	7	5	☐	0.805	Noun	Alpha
hecker	7	5	☐	0.805	Prop	Alpha
⊞ auto-gate	6	5	☐	0.804	Noun	Mixed
untouched	6	5	☐	0.804	Adj	Alpha
santorini	6	5	☐	0.804	Prop	Alpha
partly	6	5	☐	0.804	Adv	Alpha
intelligently	6	5	☐	0.804	Adv	Alpha
nine	6	5	☐	0.804	Num	Alpha
traffic	6	5	☐	0.804	Noun	Alpha
⊞ satisfy	6	5	☐	0.804	Verb	Alpha
splendid	6	5	☐	0.804	Adj	Alpha
ten	6	5	☐	0.804	Num	Alpha
⊞ yesterday	6	5	☐	0.804	Noun	Alpha
⊞ charismatic	6	5	☐	0.804	Adj	Alpha

Words, such as “hodges”, “easterly”, “tour”,“hecker” , “auto-gate” and more were appearing in only 5-6 times in the observations. It implies that those words were having very less importance and the weights were set to closer to 1.Hence those words were left from the term list.
In addition, words such as “absp” also left from the list as they did not have any meaning to build topics as it represent the space character.



Above diagram shows the number of terms in each category that were dropped or kept.Less effective 40537 terms had been dropped from the term list.



The Number of Documents by Weight plot shows the number of documents in which each term appears relative to each term’s weight. The majority of the words have rank between 35 to 700 approximately.

Generate Topics in Text Topic node

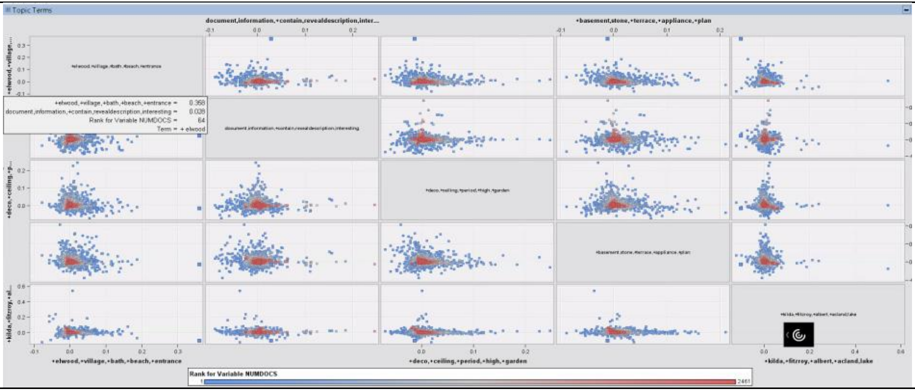
10 topics had been created using the most significant terms and it shows as below.

Topics

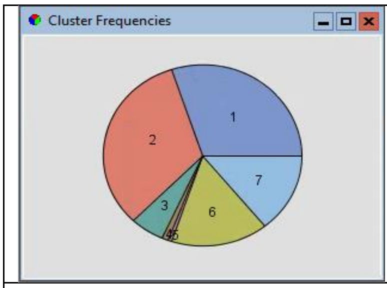
Topic	Category	Term Cutoff	Document Cutoff	Number of Terms	# Docs
+basement,stone,+terrace,+appliance,+plan	Multiple	0.027	0.119	250	462
+buyer,+property,+investor,+caulfield,monash	Multiple	0.027	0.132	271	453
+complement,+enhance,+accompany,+inviting,+appeal	Multiple	0.027	0.115	245	276
+deco,+ceiling,+period,+high,+garden	Multiple	0.027	0.104	219	421
+elwood,+village,+bath,+beach,+entrance	Multiple	0.027	0.154	204	409
+kilda,+fitzroy,+albert,+acland,lake	Multiple	0.026	0.102	168	452
+reverse-cycle,+conditioner,+price,+caulfield,air	Multiple	0.027	0.096	250	393
+security,+retreat,+carport,+cafe,+blind	Multiple	0.027	0.143	236	282
+villa,+garage,+courtyard,+unit,ducted	Multiple	0.027	0.114	236	414
document,information,+contain,revealdescription,interesting	Multiple	0.027	0.077	187	27

The combination of 5 terms which occur between numbers of documents 462 to 27 have been Identified as top 10 topics. In 5th topic, It shows that suburb names such as St Kilda, Fitzroy, albert and Acland occurs in majority of description where the house price is relatively high and it might significantly help the model to explain the variation of price.

Text Analytics in EM (Page 3)



Above chart shows how topics relate to each other. In Topic 5 and Topic 10, the term "elwood" does not have a significant relationship and it explains the same Behaviour across other relationships between topics. Observing less significant terms as such shall drop from the text filter node to improve the accuracy of text topics.



Clusters

Cluster ID	Descriptive Terms	Frequency	Percentage
1	+shop +train +home +air +room ...	846	30%
2	+floor +park +robe open +balcony ...	924	33%
3	+caulfield monash +train +shop +car ...	157	6%
4	+duty +stamp +saving +completion +desig...	38	1%
5	+accept +contain +enquiry +error +inaccur...	14	0%
6	+kilda secure +fitzroy +balcony +car ...	447	16%
7	+ceiling polished +garden +home +room ...	398	14%

Based on terms on description, 7 clusters have been generated along with cluster_SVD and cluster_prob variables against each observations as input to data which convert text data into meaningful attributes to use in the model.

Text analytic on Title variable

Terms

Term	Role	Attribute	Status	Weight	Imported Frequency	Freq	Number of Imported Documents	# Docs	Rank
+ lifestyle	...Noun	Alpha	Keep	0.375	141	143	141	143	1+
+ location	...Noun	Alpha	Keep	0.384	132	133	132	133	2+
+ apartment	...Noun	Alpha	Keep	0.400	117	117	117	117	3+
st	...Prop	Alpha	Keep	0.403	115	115	115	115	4
deco	...Prop	Alpha	Keep	0.403	115	115	115	115	4
+ living	...Prop	Alpha	Keep	0.412	102	107	102	107	6+
+ sell	...Verb	Alpha	Keep	0.419	89	104	89	102	7+
spacious	...Adj	Alpha	Keep	0.424	97	97	97	97	8
+ courtyard	...Noun	Alpha	Keep	0.424	93	97	93	97	8+
+ kilda	...Prop	Alpha	Keep	0.425	95	96	95	96	10+
+ style	...Noun	Alpha	Keep	0.439	79	86	79	86	11+
+ bright	...Adj	Alpha	Keep	0.455	69	76	69	76	12+
apartment	...Prop	Alpha	Keep	0.465	70	70	70	70	13
+ position	...Noun	Alpha	Keep	0.465	68	70	68	70	13+

Words, such as "lifestyle", "location", "apartment" were appearing in most of the titles. The word "lifestyle" is recorded 143 times in 143 observation and rank as first.

Topics

Topic	Category	Term Cutoff	Document Cutoff	Number of T...	# D...
+sell,+vendor,+price,+owner,+huge	Multiple	0.052	0.148	4	102
st,+kilda,heart,ode,kildas	Multiple	0.056	0.135	6	115
+lifestyle,+location,+stylish,living,+style	Multiple	0.06	0.124	10	251
+offer,+owner,+apartment,+dream,+beach	Multiple	0.052	0.113	1	52
deco,art,art,art,+delightful	Multiple	0.061	0.111	19	140
+apartment,+courtyard,+style,bedroom,+huge	Multiple	0.063	0.117	25	286
spacious,+stylish,+living,apartment,secure	Multiple	0.063	0.112	21	253
+location,+great,investment,+first,+home	Multiple	0.063	0.111	21	289
+big,+bright,modern,light,+courtyard	Multiple	0.06	0.106	13	136
+position,+living,+perfect,+style,perfect	Multiple	0.064	0.111	23	312

Clusters

Cluster ID	Descriptive Terms	Frequency	Percentage
1	+offer st +kilda contrac ode ...	120	4%
2	deco +lifestyle +location +apartment +living ...	1412	50%
3	lifestyle +view +live park +opportunity ...	1051	37%
4	+sell ...	241	9%

Based on terms on title, 10 text topics and 4 clusters have been generated along with cluster_SVD and cluster_prob, variables against each observations as input to data which convert text data into meaningful attributes.

Predictive Models in EM (Page 1)- Using structured data ONLY

The figure1 depicts the Predictive Model by inputting all significant structured variables and targeting the House Price for prediction. Here the Data is split in 40% training, 30% Validation and 30% test datasets. Based on interval target, Gradient Boosting, Regression, Decision Tree and Neural Network Models are mostly used models for predictions. The figure2 shows the generated decision tree with root node as a bedroom based on split selection from training dataset which is '2' which determines that there are more properties with 2 or more bedrooms. The Average squared error is more preferable as a model validation criteria considering interval target.

As shown in figure3, the neural network illustrates almost similar results for both training and validation dataset which reveals that the model has predicted correctly until price is in the range of $800000. Above this price fluctuations in the graph shows wrong predictions.

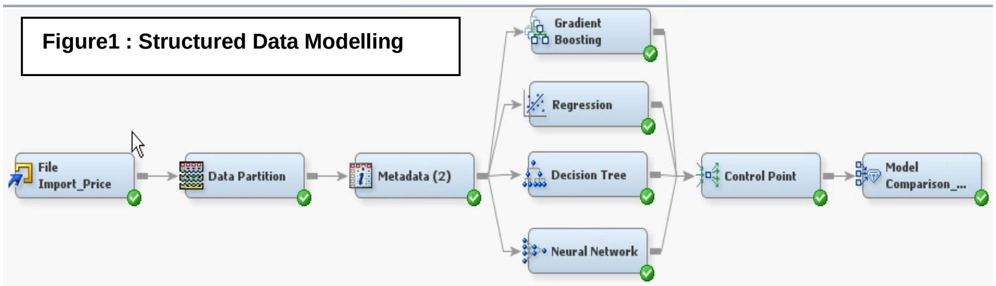


Figure1 : Structured Data Modelling

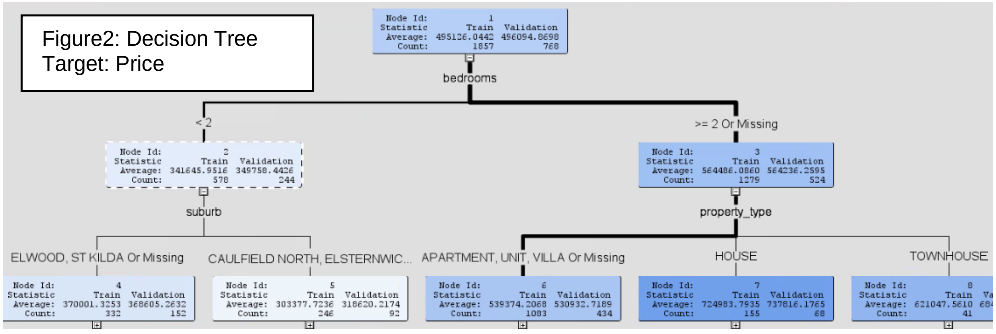


Figure2: Decision Tree
Target: Price

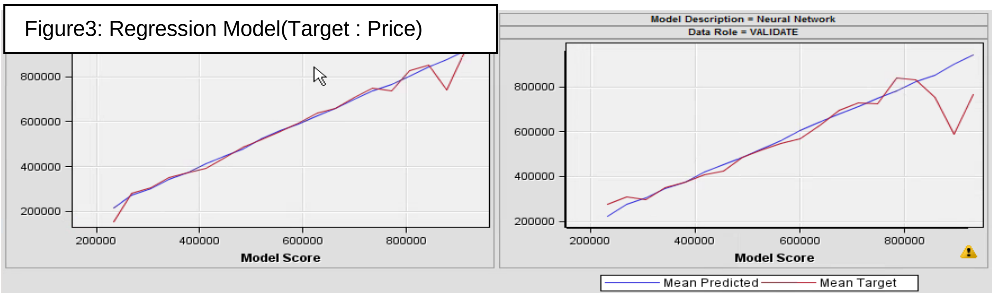


Figure3: Regression Model(Target : Price)

Figure4 depicts the fluctuations of predicted web visits and target web visits for both training and validation dataset. It can be clearly determined that there predictions are good only in between the range of 900 to 1800. Regression Model is not powerful to predict the web visits below 900 and above 1800.

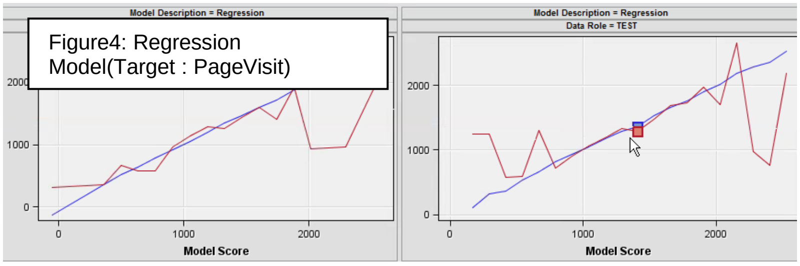


Figure4: Regression Model(Target : PageVisit)

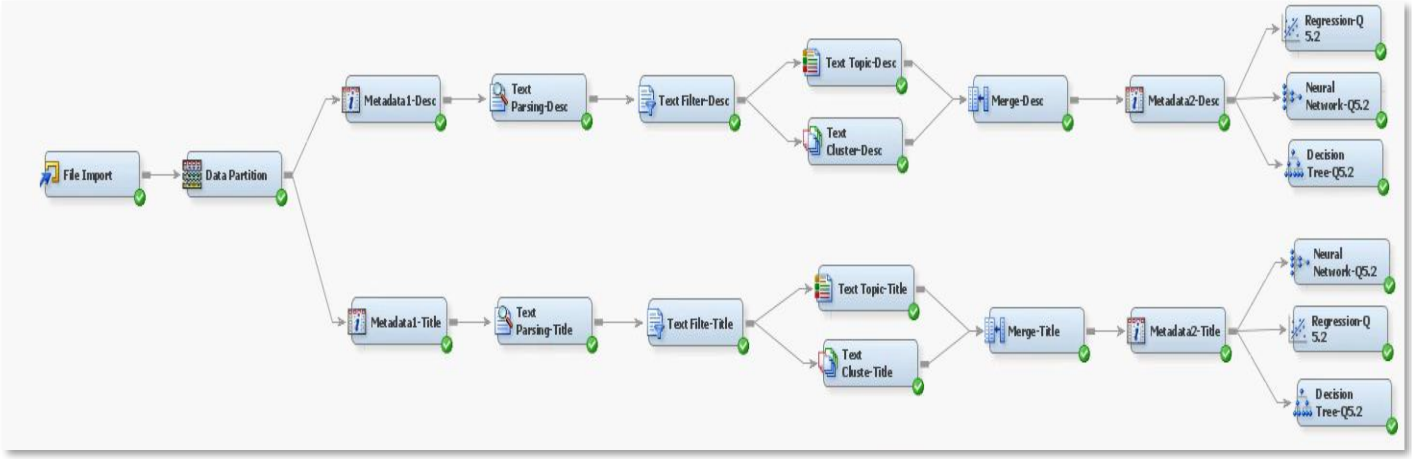
Model Description	Target	Selection Criterion: Train: Average Squared Error	Train: Root Average Squared Error	Valid: Root Average Squared Error
Neural Network	page_visits	435992.2	660.297	693.6463
Decision Tree	page_visits	441953.7	664.7959	702.0932
Regression	page_visits	447087.4	668.6459	710.2984

Model Description	Target Variable	Selection Criterion: Train: Average Squared Error	Train: Root Average Squared Error ▲	Valid: Root Average Squared Error
Neural Net...	price	8.2685E9	90931.3	96321.47
Regression	price	8.9025E9	94353.29	99285.42
Decision Tr...	price	8.9831E9	94779.04	100228.6
Gradient Bo...	price	1.21E10	109984.8	115628.8

Root Average Squared Error for training and valid dataset for all three models on page visits is as above. RMSE for Neural network for validation data is 693 which means that the actual value can be in the range between the predicted value +- 693. Similarly decision tree and regression models shows higher variation, 702 and 710 visits as RMSE.

RMSE for the Models targeting property price is as above. All models can be considered good considering due to less difference in their RMSE between each other. The Neural network model predicts the price more accuracy than Others models. For predicted value of $800000, the actual price can be in the range of $704000 - $$896000.

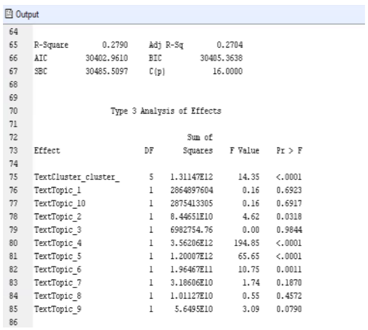
Predictive Models in EM (Page 2)- Using generated 10 text topics and clusters



Above distinct analytic models were developed to predict house price using text data as input characteristics. As the data set has two input text data characteristics(Description and title variables), separate models were developed using relevant text topic and clusters. Similarly ,distinct analytic models were developed to predict interest(page visits)

Using only text data ("description" text variable) to predict house price

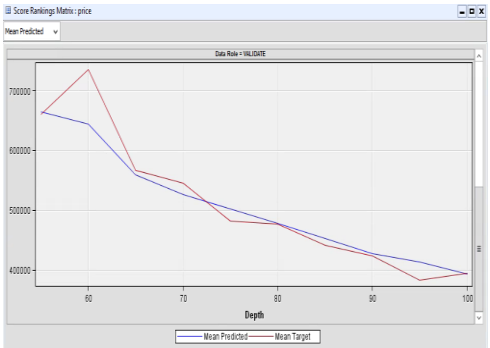
Regression Model



```
Output
64
65   R-Square        0.2790     Adj R-Sq        0.2704
66   AIC         30402.9610     BIC         30405.3638
67   SBC         30485.5097     C(p)           16.0000
68
69
70                  Type 3 Analysis of Effects
71
72                                         Sum of
73   Effect                      DF       Squares    F Value    Pr > F
74
75   TextCluster_cluster_         5    1.31147E12      14.35    <.0001
76   TextTopic_1                  1    2864897604       0.16    0.6923
77   TextTopic_10                 1    2875413305       0.16    0.6917
78   TextTopic_2                  1    8.44651E10       4.62    0.0318
79   TextTopic_3                  1    6982754.76       0.00    0.9844
80   TextTopic_4                  1    3.56206E12     194.85    <.0001
81   TextTopic_5                  1    1.20007E12      65.65    <.0001
82   TextTopic_6                  1    1.96467E11      10.75    0.0011
83   TextTopic_7                  1    3.18606E10       1.74    0.1870
84   TextTopic_8                  1    1.01127E10       0.55    0.4572
85   TextTopic_9                  1     5.6495E10       3.09    0.0790
86
```

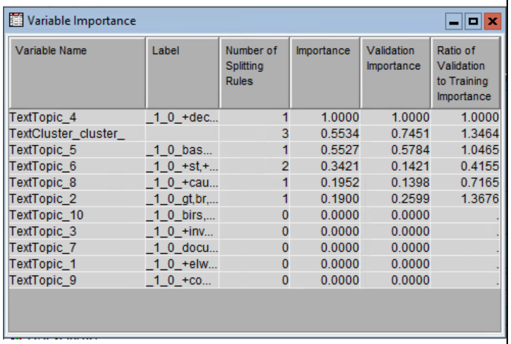
10 text topics were constructed as inputs. However, only 29% of price are explained by those inputs.

Neural Network Model



For some values of ranges the prediction is above or below of the true values.

Decision Tree Model



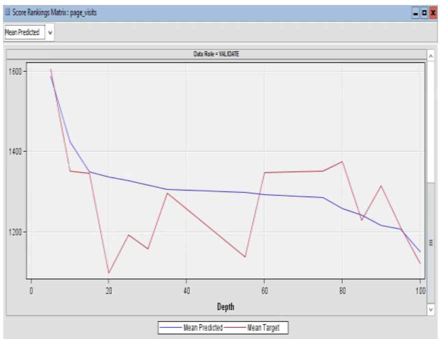
Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
TextTopic_4	_1_0_+dec...	1	1.0000	1.0000	1.0000
TextCluster_cluster_		3	0.5534	0.7451	1.3464
TextTopic_5	_1_0_bas...	1	0.5527	0.5784	1.0465
TextTopic_6	_1_0_+st,+...	2	0.3421	0.1421	0.4155
TextTopic_8	_1_0_+cau...	1	0.1952	0.1398	0.7165
TextTopic_2	_1_0_gt,br,...	1	0.1900	0.2599	1.3676
TextTopic_10	_1_0_birs,...	0	0.0000	0.0000	
TextTopic_3	_1_0_+inv...	0	0.0000	0.0000	
TextTopic_7	_1_0_docu...	0	0.0000	0.0000	
TextTopic_1	_1_0_+elw...	0	0.0000	0.0000	
TextTopic_9	_1_0_+co...	0	0.0000	0.0000	

Topic_4 was identified as the most important variable as the first node.

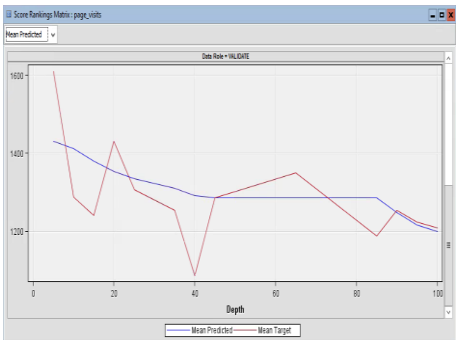
Using only text data ("title" text variable) to predict interest (web visits)

Regression Model



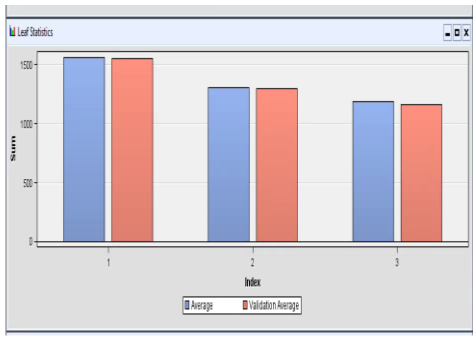
For most values of ranges the prediction is far above of the true values

Neural Network Model



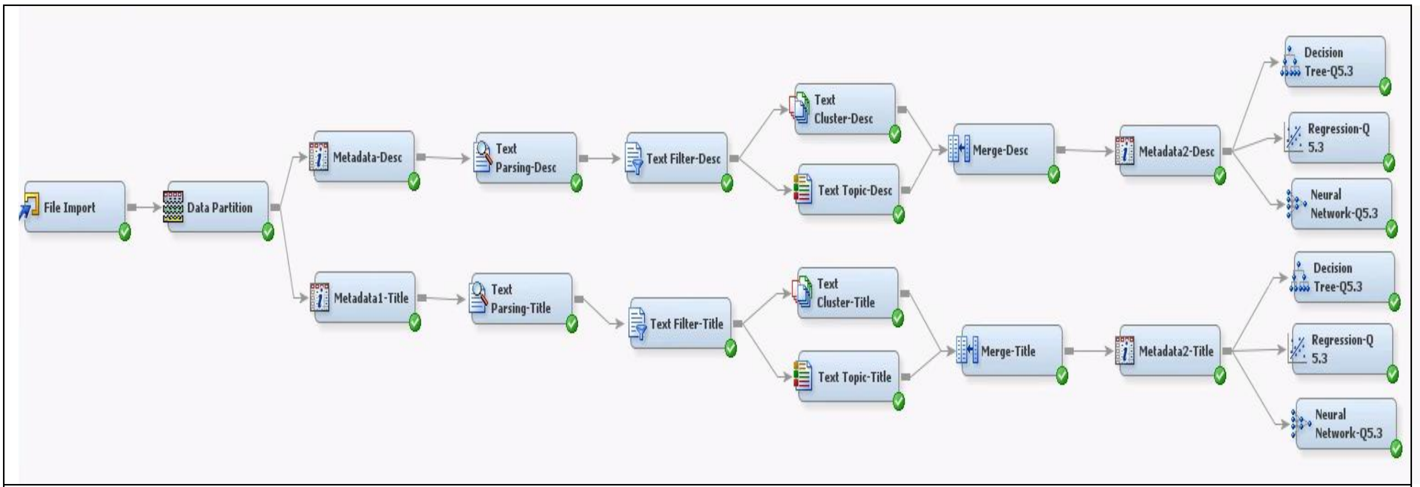
For most values of ranges the prediction is far above of the true values.

Decision Tree Model



AS similar increase in bar heights would be seen in the validation data, the model is reliable.

Predictive Models in EM (Page 3)-Using generated 10 text topics and 5 structured data



Above distinct analytic models were developed to predict house price using **both structure and text data as input characteristics**. As the data set has two input text data characteristics(Description and title variables), separate models were developed using relevant text topic and clusters. Similarly, distinct analytic models were developed to predict interest(page visits) using both structure and text data.

Using only text data ("description" text variable) to predict house price

Regression Model

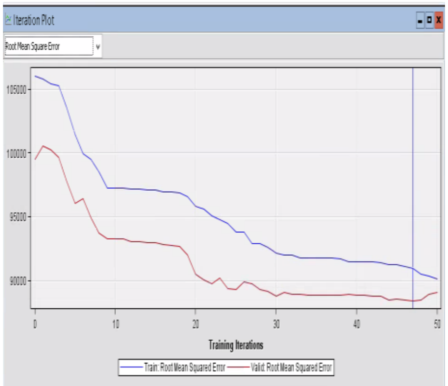
R-Square	0.7139	Adj R-Sq	0.7008
AIC	30739.9879	BIC	30747.5736
SBC	31052.3705	C(p)	60.0000

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
TextCluster_cluster_	6	9.93503E10	2.16	0.0445
TextTopic_1	1	2.32153E10	3.03	0.0822
TextTopic_10	1	4.8639E10	6.34	0.0119
TextTopic_2	1	1.48876E10	1.94	0.1638
TextTopic_3	1	8343847597	1.09	0.2972
TextTopic_4	1	9.88931E11	128.93	<.0001
TextTopic_5	1	1.99538E11	26.01	<.0001
TextTopic_6	1	3249539277	0.42	0.5152
TextTopic_7	1	1.37901E11	17.98	<.0001
TextTopic_8	1	1.94075E10	2.53	0.1119
TextTopic_9	1	2.56231E11	33.40	<.0001
bathrooms	1	5.45365E11	71.10	<.0001
bedrooms	1	5.27448E12	687.63	<.0001
car_spaces	4	3.9244E11	12.79	<.0001
property_type	6	1.76384E12	38.32	<.0001
suburb	31	1.91221E12	8.04	<.0001

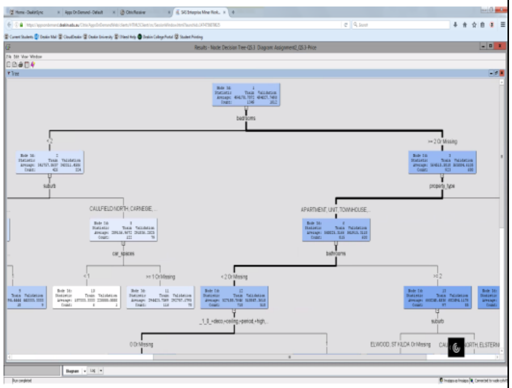
10 text topics and 5 structured variables were used as inputs.70% of price variation are explained by those inputs.

Neural Network Model



Validation data set has always less RMSE than training dada set.

Decision Tree Model



Bedrooms variable is identified as the root node. Distribution for training and validation data in root node has similar statistics.

Using only text data ("title" text variable) to predict interest (web visits)

Regression Model

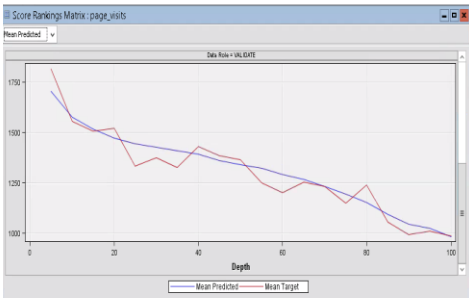
R-Square	0.0776	Adj R-Sq	0.0570
AIC	35833.8145	BIC	35838.5872
SBC	36194.7620	C(p)	61.0000

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
TextCluster2_cluster_	4	1172099.36	0.64	0.6352
TextTopic2_1	1	677277.658	1.48	0.2246
TextTopic2_10	1	298249.834	0.65	0.4203
TextTopic2_2	1	617955.453	1.35	0.2461
TextTopic2_3	1	125190.102	0.27	0.6016
TextTopic2_4	1	26204.9173	0.06	0.8112
TextTopic2_5	1	625705.823	1.36	0.2431
TextTopic2_6	1	605878.935	1.32	0.2507
TextTopic2_7	1	276288.137	0.60	0.4380
TextTopic2_8	1	3569391.49	7.78	0.0053
TextTopic2_9	1	17024.4567	0.04	0.8473
bathrooms	2	1704250.82	1.86	0.1565
bedrooms	2	3093576.51	3.37	0.0346
car_spaces	4	3376793.57	1.84	0.1186
property_type	7	23657915.5	7.36	<.0001
suburb	31	55186481.5	3.88	<.0001

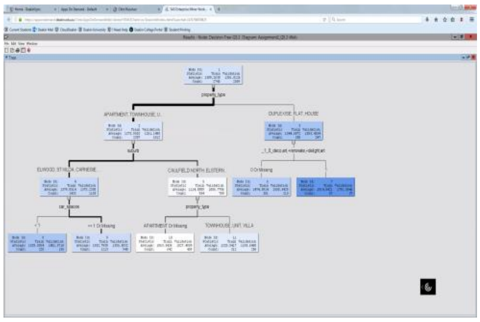
variation on web visits do not explain significantly by input variables.

Neural Network Model



For some values of ranges the prediction is above or below of the true values.

Decision Tree Model



Property type is identified as the root node followed by suburb and text topic variable.

Model Comparison and Ensemble Models (Page 1)

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Valid: Root Average Squared Error ▼	Test: Root Average Squared Error
	Reg5	Reg5	Regression (5)	price	97947.31	98736.14
	Tree5	Tree5	Decision Tree (5)	price	97038.02	102661.1
Y	Neural5	Neural5	Neural Network (5)	price	96063.18	98273.71

The Fit-statistics explains the Root Average Square Error against each model to predict price, sung structured data ONLY as input variables .It is clear that NN predicts with more accuracy as it has the lowest Root Average Squared error of $96063 for validation data.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Valid: Root Average Squared Error	Test: Root Average Squared Error
Y	Reg	Reg	Regression-Q5.2	price	136602.4	137525.6
	Tree	Tree	Decision Tree-Q5.2	price	137956.2	139152.3
	Neural	Neural	Neural Network-Q5.2	price	138089.4	138391.5

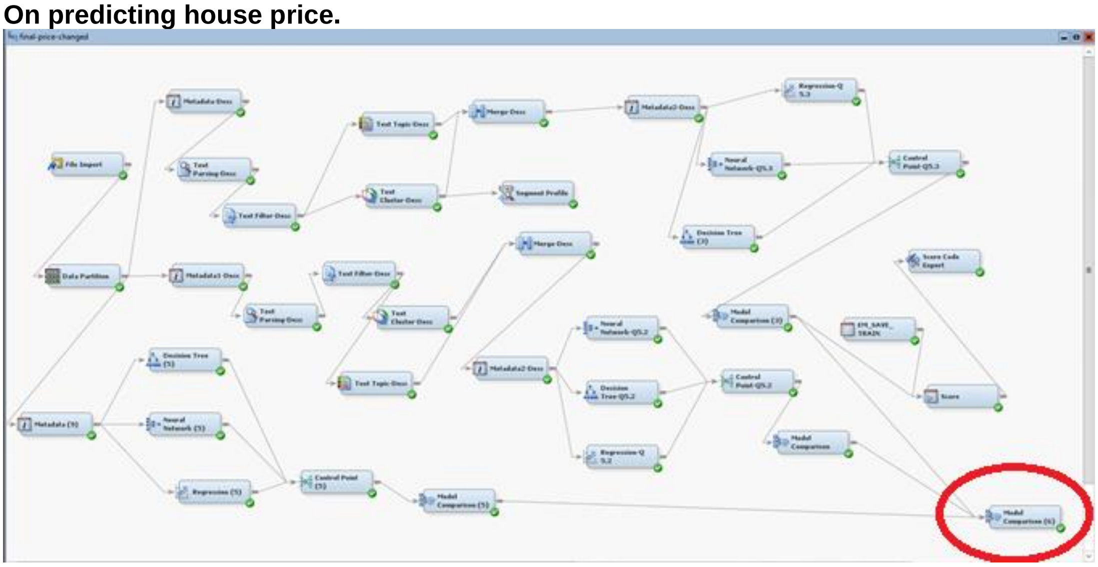
The Fit-statistics explains the Root Average Square Error against each model to predict price, sung text data ONLY as input variables.
According to the statistics, that regression model predicts with more accuracy as it has the lowest Root Average Squared error of $136602 for validation data.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Valid: Root Average Squared Error ▲	Test: Root Average Squared Error
Y	Reg3	Reg3	Regression-Q5.3	price	82980.25	83988.17
	Neural3	Neural3	Neural Network-Q5.3	price	86137.25	91119.88
	Tree3	Tree3	Decision Tree (3)	price	89915.04	97181.31

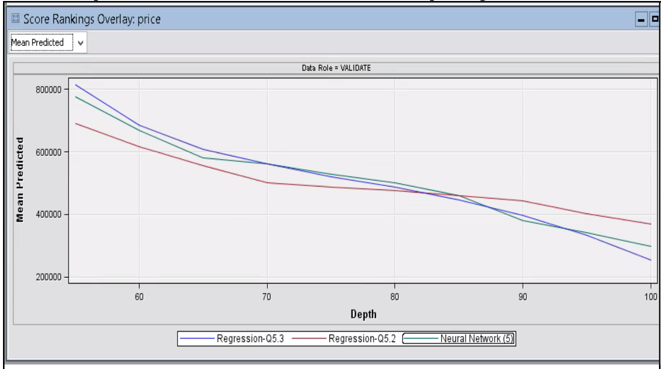
The Fit-statistics explains the Root Average Square Error against each model to predict price, sung structured and text data as input variables.
I emphasis, that regression model predicts with more accuracy as it has the lowest Root Average Squared error of $82980 for validation data over Neural network and Decision tree models. This accuracy is around +- $82980 compared to other 2 models.

Model Comparison between models using structured data only, text data only and both structured and text data On predicting house price.



Model Comparaison and Ensemble Models (Page 2)

Champion Predictive Model: Property Price



Fit Statistics

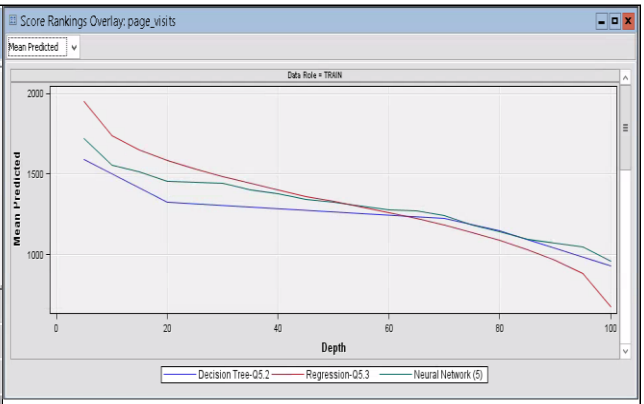
Selected Model	Predecessor Node	Model Node	Model Description	Target	Valid: Root Average Squared Error	Test: Root Average Squared Error
	MdlComp5	Neural5	Neural Network (5)	price	96063.18	98273.
	MdlComp	Reg	Regression-Q5.2	price	136602.4	137525
Y	MdlComp3	Reg3	Regression-Q5.3	price	82980.25	83988.

Comparing all models upon different inputs, the best predictive model to predict price is the regression model which use both structure and text data as inputs. As inputs the model use property type, bedrooms, bathrooms, car space and suburb variables and text variable data generated from description variable. On F statistics for model comparison, it shows the best accuracy on regression model as it has less RASE on validation data set of $82980. The property price prediction on models between $500000 shows more similar figures according to overlay graph. In addition, property price prediction more than $500000(approximately) regression model predicts more price than other two models while price prediction less than around $500000 regression model predicts less price over other two models.

Champion Predictive Model: Page visits

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Valid: Root Average Squared Error	Test: Root Average Squared Error
	MdlComp	Tree	Decision Tree-Q5.2	page_visits	690.5906	697.1617
	MdlComp3	Reg3	Regression-Q5.3	page_visits	677.3529	680.3143
Y	MdlComp5	Neural5	Neural Network (5)	page_visits	674.9504	689.3899



Comparing all models upon different inputs, the best predictive model to predict number of page visits to determine the public interests is the Neural Network Which use both structure and text data as inputs. As inputs the model use property type, bedrooms, bathrooms, car space and suburb variables and text variable data generated from description variable. On F statistics for model comparison, it shows the best accuracy on Neural Network as it has less RASE on validation data set of 674 page visits. The graph shows that the number of page visits between1000 to 1300 can be predicted mostly similar with all three models but out of this range Neural Network predicts with more accuracy and hence it is a selected model for prediction of page visits.

Building model to predict target variables, both house price and page visits gives more accuracy on using both text and structured data, implies that text data help to explain the variation on target variable. In summary, we convert these text data (description, Title variable) into structured data using text analytics which actually contributes significantly on prediction model.

References
"Property Data - Consumer Affairs Victoria". *Consumer.vic.gov.au*. N.p., 2016. Web. 26 Sept. 2016.