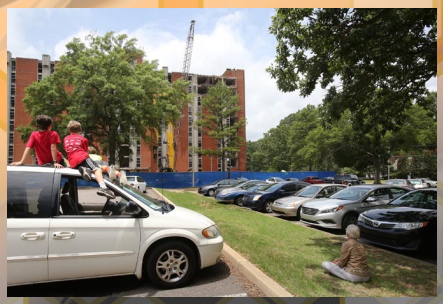
Blight Property



1

Agenda

- Background
- Introduction and goal
- Data source and description
- Key challenges
- Descriptive analysis(tableau)
- Predictive analysis(SAS EM)
- Key Learnings

2

Background of the blight property

Blight: abandoned and/or vacant property, either building or home, that is in an unacceptable condition of disrepair. It is inhabitable and has little to no economic value.

Background:

- US is facing the problems of empty neighborhoods across the country. We can see similar problems in Memphis, TN, and the causes for this are: Poor planning, Poverty and Sub Urbanization are the causes of the Urban Blight. The problems are caused because of – Unemployment, depopulation problems, promote crime, loss of deindustrialization. Blight affected areas are characterized by dirty, cloistered, tumbledown buildings and inhospitable living.

Effects of such properties

- Danger to the public
- Increase in negative incidents
- Negative effect on surrounding property values
- Reduces property tax revenue

3

Business objective and project goal

- Business objective:

Identify properties that are possible to become blighted. This information can help local community and government plan in advance and minimize the bad effects of a blighted property/area.

- Project goal

Identify the data that are more related to to blight property, characterize their relationships and predict the possibility that a property turn into a blight property

- Approaches
  - Use demolition fees as an indicator of blight properties.
  - Explore relevant data (i.e., crime, sales, blight rating, property tax etc.) with tableau, and try to find the data related to blight property and explore their relationships.
  - Explore the data with SAS EM to further characterize the relationships between the relevant data and the target and predict the possibilities.

4

Data source and description

Datasets	Content of each data set
Windshield	The thorough housing information , condition details of all the Memphis and Shelby property, and Memphis and Shelby county regulation and code enforcement, ~150,000 parcel ID. The data set was not cleaned
Tax aggregation	32 statement records of aggregated (sum) of property tax information over the period of 2010-2016, 46,751 parcel ID, data collected from 2010 to 2016.
MLGW aggregation	count how many times MLGW cut off the supply to one property. Interval (discrete) data with data range [0, 2], two types of data recorded (gas cut and electricity cut) per year for 15,678 parcel ID
Sales aggregation	count how many times sale happens each year. Interval (discrete) data with data range [0, 27], one sale record of each year for 127,609 parcel ID
Crime aggregation	count how many crimes happen on each property. Interval (discrete) data with data range [0, 447], three types of crimes(violent, nonviolent, other) document for each year; 98,366 data point

- Data processing approach and flow for further analysis:
  - Randomly sampling parcel IDs in the tax aggregation data and compare the tax value with the information on the Shelby county trustee website
  - Select the tax aggregation data from a year that has the most data points documented
  - Clean up all types of errors in the Windshield data
  - Convert Parcel IDs in the above five datasets into the same format.
  - Inner joined all the dataset and create a dummy variable for demolition fees.
- Target variable: demolition dummy variable
- Input variables for SAS EM:
  - 1st round, all the variables from the above data set except for unary data, ID, zip code and data missing over 50%
  - 2nd and the later round, all the variables that have the P value lower than 0.6 in the regression model statistics

5

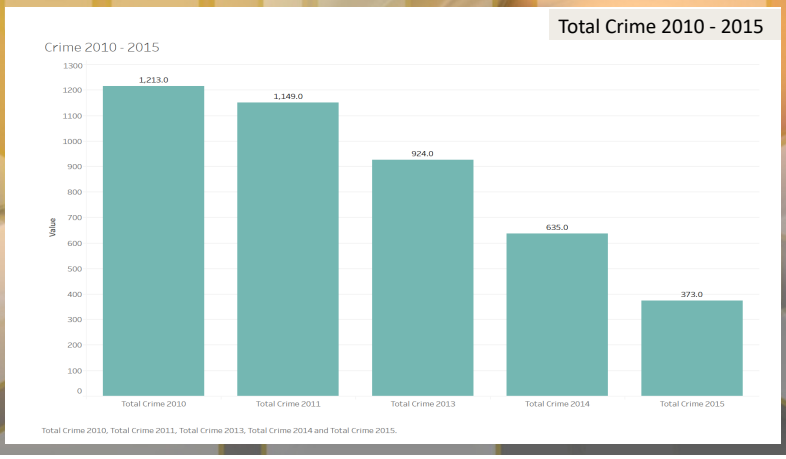
Key Challenges

- Data challenges:
  - Data cleaning, the windshield data provide useful details of housing conditions but it is very messy. It took about 8 hours to clean it to a useful/informative level.
  - The inaccuracy of some given datasets requires to examine the data using higher authority data source, which was not expected when the project started.
  - Need domain knowledge of housing, and Shelby county regulation to understand the meaning and value of related variables and manipulate the input without biasing the modeling process.
  - Too many variables for EM regression models and missing values in a high dimension dataset
- Data techniques:
  - Descriptive analysis with Tableau
  - Joint or merge datasets, Query and computing in SAS EG
  - Decision trees, regression and neural network modeling with SAS EM

6

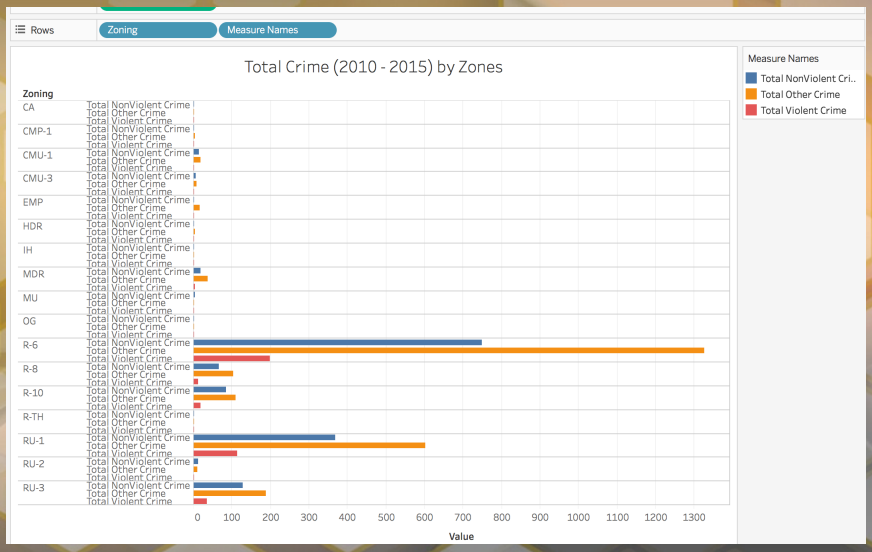
Descriptive Analytics

- Analyze datasets to get basic understanding of simple statistics
- Define the most affected zoning districts
- Analyze and explore relationship between Crime, Sales, Gas/Power cuts, property tax and demolition fees.



7

Violent, Non Violent, Other Crime by Zones



8

Crime (2010 – 2015) by Zones

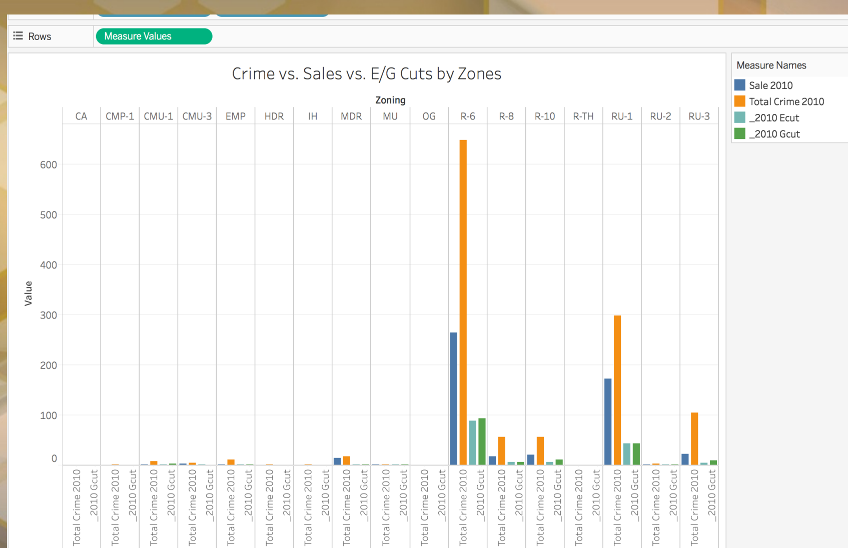
Rows: Zoning, Measure Values

Crime 2010 - 2015

Zoning	Total Crime 2010	Total Crime 2011	Total Crime 2013	Total Crime 2014	Total Crime 2015
CA	0.0	3.0	0.0	0.0	0.0
CMP-1	1.0	1.0	0.0	4.0	2.0
CMU-1	8.0	7.0	5.0	9.0	6.0
CMU-3	5.0	8.0	1.0	2.0	1.0
EMP	11.0	6.0	2.0	2.0	2.0
HDR	1.0	2.0	4.0	0.0	0.0
IH	1.0	0.0	0.0	0.0	0.0
MDR	17.0	19.0	8.0	10.0	6.0
MU	1.0	2.0	0.0	2.0	0.0
OG	0.0	1.0	0.0	1.0	1.0
R-6	648.0	597.0	479.0	376.0	173.0
R-8	56.0	43.0	32.0	31.0	21.0
R-10	56.0	58.0	51.0	26.0	22.0
R-TH	0.0	1.0	0.0	0.0	0.0
RU-1	299.0	299.0	231.0	138.0	119.0
RU-2	4.0	5.0	6.0	5.0	3.0
RU-3	105.0	97.0	105.0	29.0	17.0

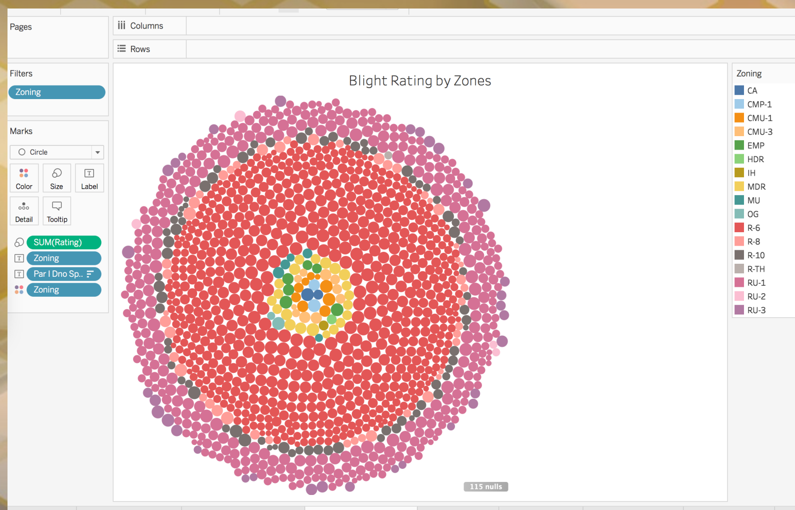
9

Gas/Power Cuts by Zones



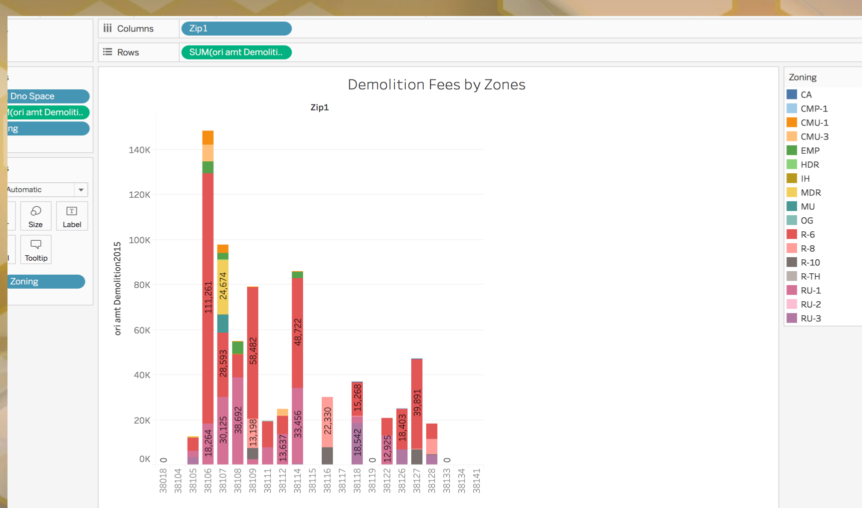
10

Blight Rating by Zones



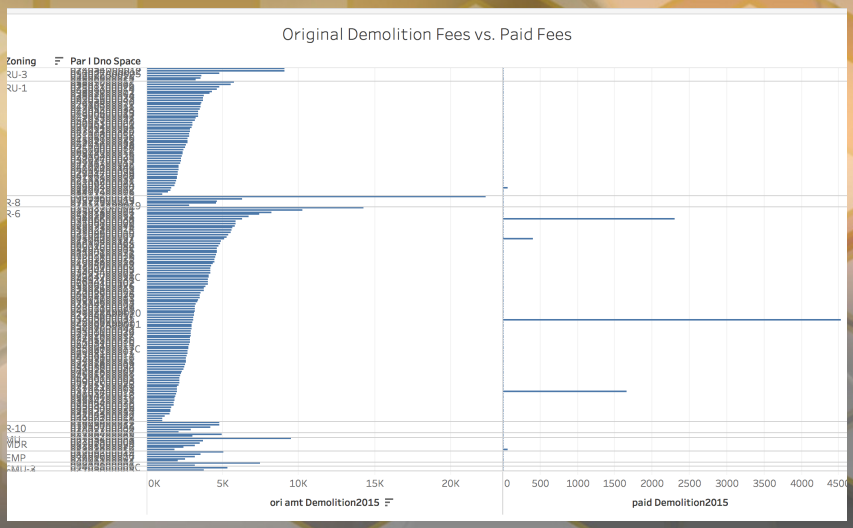
11

Demolition Fees by Zones



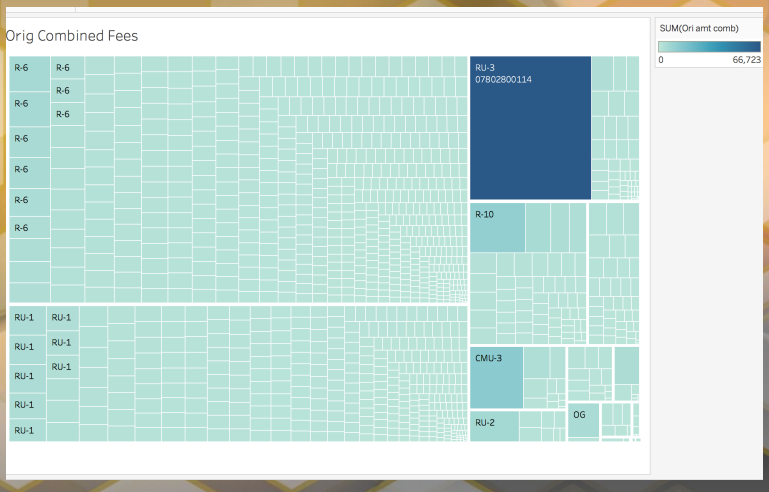
12

Original Demolition Fees vs. Paid Fees



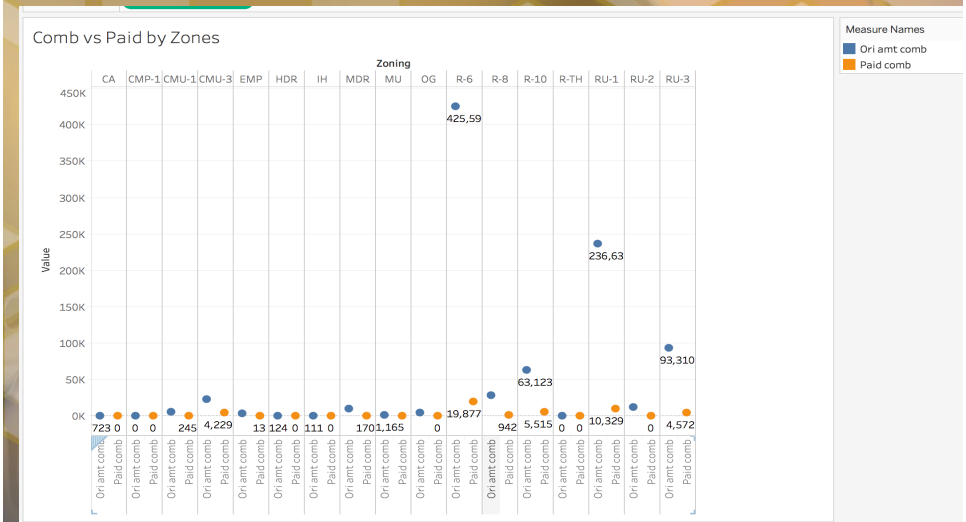
13

Combined Fees Due



14

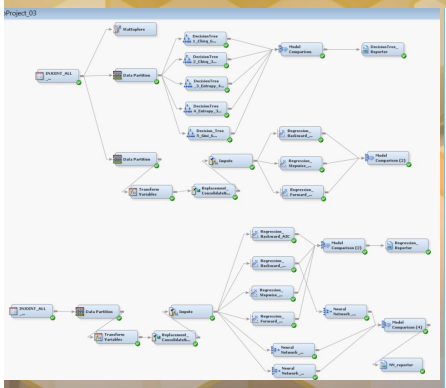
Combined Fees vs. Paid Fees

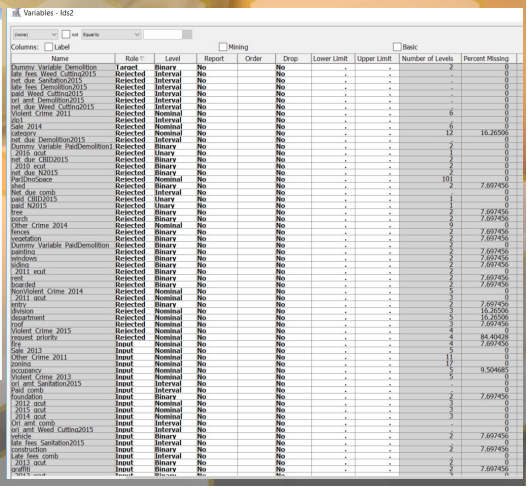


15

Predictive Analytics

The Diagram of decision trees, regression and NN models
An iterative selection approach to select independent variables

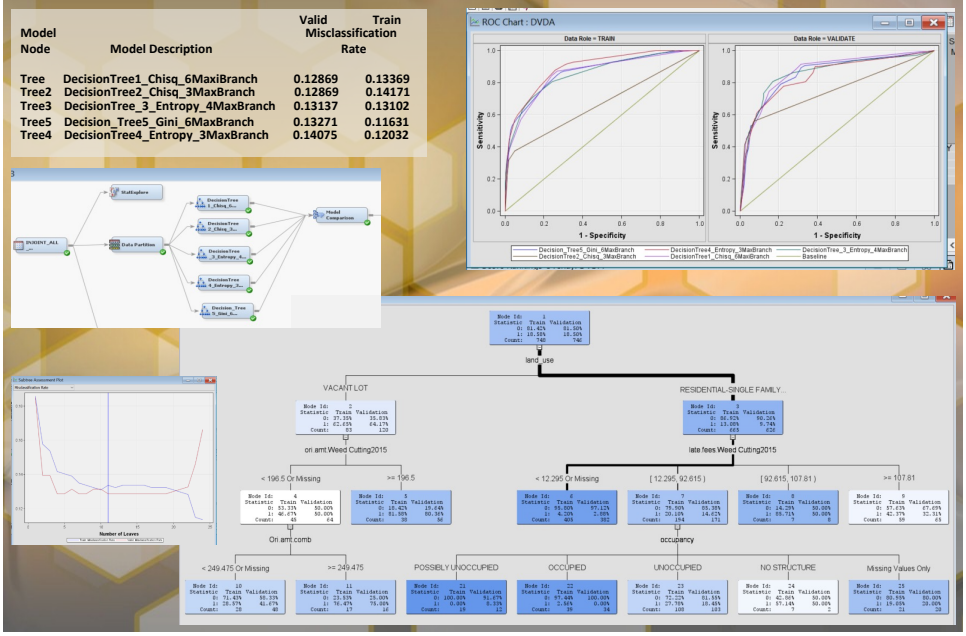




16

Decision tree model comparison and statistics

Model Node	Model Description	Valid Misclassification Rate	Train Misclassification Rate
Tree	DecisionTree1_Chisq_6MaxiBranch	0.12869	0.13369
Tree2	DecisionTree2_Chisq_3MaxBranch	0.12869	0.14171
Tree3	DecisionTree_3_Entropy_4MaxBranch	0.13137	0.13102
Tree5	Decision_Tree5_Gini_6MaxBranch	0.13271	0.11631
Tree4	DecisionTree4_Entropy_3MaxBranch	0.14075	0.12032

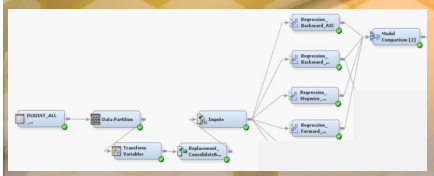


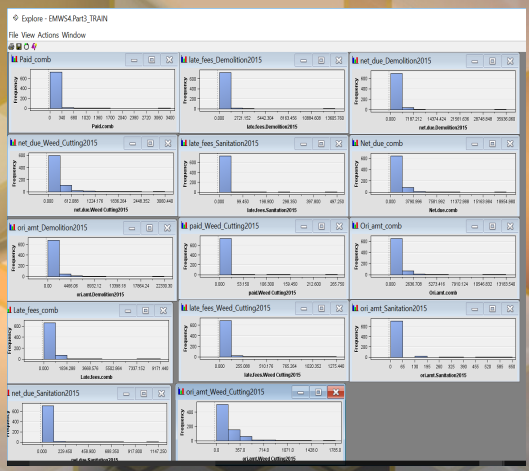
17

Data transformation and replacement for regression models

Replacement Editor-WORK.OUTCLASS

Variable	Formatted Value	Replacement Value	Frequency Count	Type
request_type	CE-Weeds Occupied Property	Weeds	63C	
request_type	CW-Weeds Vacant Houses	Weeds	9C	
request_type	CW-Weeds Vacant Lots	Weeds	2C	
zoning	RU-1	RU	196C	
zoning	RU-3	RU	30C	
zoning	RU-2	RU	4C	
zoning	R-6	R	402C	
zoning	R-10	R	42C	
zoning	R-8	R	37C	
request_type	CE-Code Miscellaneous	Miscellaneous	226C	
request_type	SC-Mayors Action Center(MIS)	Miscellaneous	1C	
zoning	CMU-1	CMU	8C	
zoning	CMU-3	CMU	6C	
zoning	CMP-1	CMP	2C	
request_type		_UNKNOWN_	121C	
request_status		_UNKNOWN_	121C	
occupancy		_UNKNOWN_	72C	
land_use		_UNKNOWN_	79C	
zoning	_UNKNOWN_	_DEFAULT_	1C	





References for variable level replacement:

The Memphis and Shelby county zoning code and subdivision regulation
https://shelbycountytn.gov/DocumentCenter/View/27740

The Memphis and Shelby county major code violation
https://www.shelbycountytn.gov/2138/Major-Code-Violations

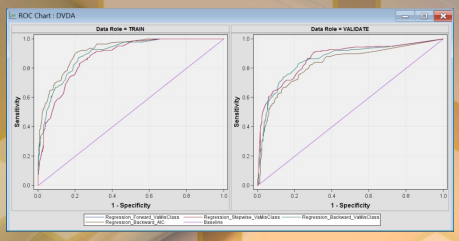
18

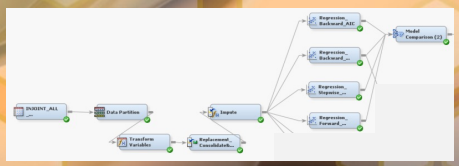
Regression model comparison and statistics

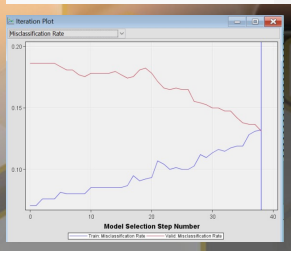
Model Node	Model Description	Valid Misclassification Rate	Train Misclassification Rate
Reg6	Regression_Stepwise_VaMisClass	0.11796	0.14305
Reg7	Regression_Forward_VaMisClass	0.11796	0.14305
Reg5	Regression_Backward_VaMisClass	0.13137	0.13235
Reg4	Regression_Backward_AIC	0.14745	0.11765

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
IMP_construction	1	11.8406	0.0006
IMP_fire	3	7.7533	0.0514
IMP_graffiti	1	10.8379	0.0010
IMP_vehicle	1	2.6360	0.1045
LOG_Late_fees_comb	1	13.3722	0.0003
LOG_Ori_amt_comb	1	10.3694	0.0013
LOG_Paid_comb	1	4.0125	0.0452
LOG_late_fees_Sanitation2015	1	3.1751	0.0748
LOG_ori_amt_Sanitation2015	1	3.1819	0.0745
LOG_ori_amt_Weed_Cutting2015	1	48.5437	<.0001
NonViolent_Crime_2010	3	1515.0422	<.0001
NonViolent_Crime_2011	4	9.8803	0.0425
NonViolent_Crime_2013	3	418.8311	<.0001
Other_Crime_2011	6	9.7427	0.1359
Other_Crime_2013	5	526.8762	<.0001
REP_occupancy	5	61.1989	<.0001
REP_request_status	6	30.6492	<.0001
Sale_2010	4	10.1077	0.0387
Sale_2013	3	384.1240	<.0001
Sale_2016	2	167.1718	<.0001
Violent_Crime_2013	1	163.7942	<.0001
_2012_gcut	1	3.3342	0.0679
_2015_gcut	1	54.3618	<.0001





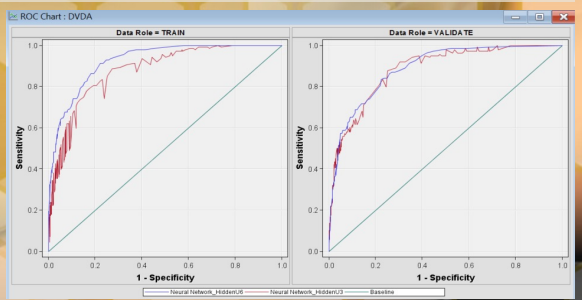


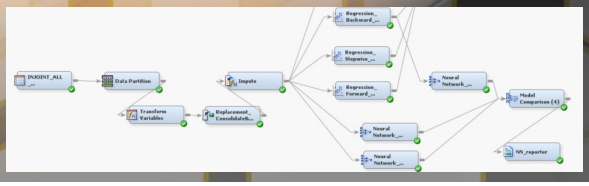
19

Neural network model comparison

Fit Statistics
Model Selection based on Valid: Misclassification Rate

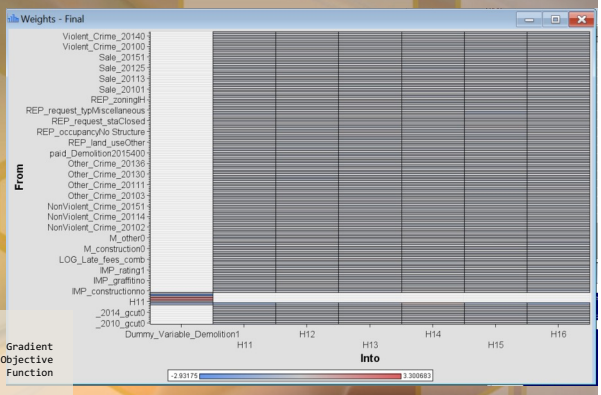
Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Neural3	Neural Network_HiddenU6	0.11930	0.08120	0.11230	0.09639
Neural2	Neural Network_HiddenU3	0.12332	0.11348	0.14572	0.10212
Neural	Neural Network_HiddenU3	0.12601	0.09640	0.13369	0.09827





20

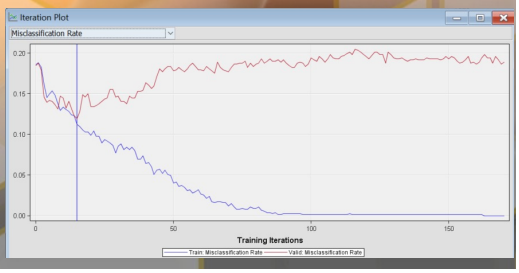
Neural network model weights and parameters



Optimization Results (1009 parameter estimates)

Parameter Estimates

N	Parameter	Estimate	Gradient Objective Function
1	LOG_Late_fees_comb_H11	0.375179	-0.000015183
2	LOG_Ori_amt_comb_H11	0.546825	-0.000017382
3	LOG_Paid_comb_H11	-0.001867	-0.000003554
4	LOG_late_fees_Sanitation2015_H11	0.431722	-0.000001962
5	LOG_ori_amt_Sanitation2015_H11	0.313230	-0.000002099
6	LOG_ori_amt_Weed_Cutting2015_H11	0.015622	0.000007933
7	LOG_Late_fees_comb_H12	1.022528	0.001136
8	LOG_Ori_amt_comb_H12	-0.970118	0.001020
9	LOG_Paid_comb_H12	-0.422006	0.000215
10	LOG_late_fees_Sanitation2015_H12	-0.787159	0.000030951
11	LOG_ori_amt_Sanitation2015_H12	0.649889	0.000014891
12	LOG_ori_amt_Weed_Cutting2015_H12	0.268840	-0.000654
13	LOG_Late_fees_comb_H13	1.087307	0.000350
14	LOG_Ori_amt_comb_H13	-0.409573	0.000466
15	LOG_Paid_comb_H13	-0.118317	-0.000117
16	LOG_late_fees_Sanitation2015_H13	-0.275618	-0.001322
17	LOG_ori_amt_Sanitation2015_H13	-0.046507	-0.000981
18	LOG_ori_amt_Weed_Cutting2015_H13	1.505582	0.000233
19	LOG_Late_fees_comb_H14	0.282938	0.000040119



21

Key learning

- Never assume the datasets are accurate or clean from any source without examining them.
- The Identification of the related variables and reduction to the unnecessary data complexity will help the SAS EM perform better---- it is not a good idea to shovel the data into EM and rely on it to perform well.
- The output log file provides a thorough documentation with statistic calculation, model parameters and iteration steps, which are helpful to trouble-shooting model optimization. Therefore, it is worth to learn how to read the file and understand certain amount of those statistics .

22

23