Exploration of Machine Learning Approaches to Predict Pavement Performance

Final Report
March 2018

Sponsored by
Midwest Transportation Center
U.S. Department of Transportation
Office of the Assistant Secretary for Research and Technology

MIDWEST TRANSPORTATION CENTER
MTC

IOWA STATE UNIVERSITY
Institute for Transportation

About MTC

The Midwest Transportation Center (MTC) is a regional University Transportation Center (UTC) sponsored by the U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology (USDOT/OST-R). The mission of the UTC program is to advance U.S. technology and expertise in the many disciplines comprising transportation through the mechanisms of education, research, and technology transfer at university-based centers of excellence. Iowa State University, through its Institute for Transportation (InTrans), is the MTC lead institution.

About InTrans

The mission of the Institute for Transportation (InTrans) at Iowa State University is to develop and implement innovative methods, materials, and technologies for improving transportation efficiency, safety, reliability, and sustainability while improving the learning environment of students, faculty, and staff in transportation-related fields.

ISU Non-Discrimination Statement

Notice

Quality Assurance Statement

Technical Report Documentation

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Exploration of Machine Learning Approaches to Predict Pavement Performance		**5. Report Date** March 2018 **6. Performing Organization Code**
7. Author(s) Ravi Nath, William Duckworth, Cindy Corritore, Nalini Govindarajulu, and Victor Ekpoke		**8. Performing Organization Report No.**
9. Performing Organization Name and Address Department of Business Intelligence & Analytics Heider College of Business Creighton University 2500 California Plaza Omaha, Nebraska 68178		**10. Work Unit No. (TRAIS)** **11. Contract or Grant No.** Part of DTRT13-G-UTC37
12. Sponsoring Organization Name and Address Midwest Transportation Center 2711 S. Loop Drive, Suite 4700 Ames, IA 50010-8664	U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology 1200 New Jersey Avenue, SE Washington, DC 20590	**13. Type of Report and Period Covered** Final Report **14. Sponsoring Agency Code**
15. Supplementary Notes Visit www.intrans.iastate.edu for color pdfs of this and other research reports.		

16. Abstract

Machine learning (ML) techniques were used to model and predict pavement condition index (PCI) for various pavement types using a variety of input variables. The primary objective of this research was to develop and assess PCI predictive models for the years 2014 and 2015 based upon the 2013 PCI values and other road characteristics during calendar year 2013. Clearly, if a road segment was resurfaced during 2014 or 2015, then this information was expected to profoundly affect the PCI for 2015. Data collected by the Iowa Department of Transportation (DOT) regarding road conditions across the state of Iowa were used to model PCI. IBM's Watson Analytics was utilized as a ML tool to perform the analysis. The analysis shows that ML is a viable approach to modelling PCI for various pavement types and that it is possible to predict future PCI from past PCI values, which thus eliminates the need to measure PCI for road segments on a yearly basis. This approach also has an advantage over multiple linear regression models in that it automatically accounts for nonlinear relationships.

17. Key Words		18. Distribution Statement	
analytics—IBM Watson Analytics—machine learning—pavement condition index—prediction—regression		No restrictions.	
19. Security Classification (of this report) Unclassified.	**20. Security Classification (of this page)** Unclassified.	**21. No. of Pages** 38	**22. Price** NA

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized**

EXPLORATION OF MACHINE LEARNING APPROACHES TO PREDICT PAVEMENT PERFORMANCE

Final Report
March 2018

Principal Investigator
Ravi Nath, Professor and Chair
Department of Business Intelligence & Analytics, Heider College of Business, Creighton University

Co-Principal Investigators
William Duckworth, Associate Professor
Cindy Corritore, Professor
Nalini Govindarajulu, Associate Professor
Department of Business Intelligence & Analytics, Heider College of Business, Creighton University

Research Assistant
Victor Ekpoke

Sponsored by
Midwest Transportation Center and
U.S. Department of Transportation
Office of the Assistant Secretary for Research and Technology

A report from
Institute for Transportation
Iowa State University
2711 South Loop Drive, Suite 4700
Ames, IA 50010-8664
Phone: 515-294-8103 / Fax: 515-294-0467
www.intrans.iastate.edu

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The authors would like to thank the Midwest Transportation Center and the U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology for sponsoring this research.

1. MODELING PAVEMENT CONDITION INDEX (PCI)

Data collected by the Iowa Department of Transportation (DOT) regarding road conditions across the state of Iowa are used to model the pavement condition index (PCI). Data consist of PCI values for the calendar years 2013 (PCI_2013), 2014 (PCI_2014), and 2015 (PCI_2015) and indicators showing whether a road segment is resurfaced either during the year 2014 or 2015. In addition, data for various road characteristics and measures for 2013 are considered. Data are available for a total of nearly 4,000 road segments.

The primary objective of this investigation was to develop and assess PCI predictive models for 2014 and 2015 based upon the 2013 PCI values and other road characteristics and measures captured during the calendar year 2013. Clearly, if a road segment was resurfaced during 2014 or 2015, then this information was also incorporated in the predictive model.

One expects PCI values to vary according to the type of materials used in constructing the pavement. Therefore, a separate analysis was conducted for each pavement type. We considered three pavement types for which sufficient data were available. These pavement types, along with the number of observations (n) and percentages, are shown in Table 1.

Table 1. Pavement types

Pavement Type	n	%
Portland Cement	1,251	34.86
Composite	1,876	52.27
Asphalt Cement	462	12.87
Total	3,589	100%

Table 2 provides the names and descriptions of the 21 relevant variables used in the analysis. The table also indicates whether a variable serves as an input variable or a target variable, as well as whether the variable is continuous or binary. Further, Tables 3 and 4 show the mean and standard deviation for each variable and are categorized by pavement types. All analyses were completed using IBM Watson Analytics software, which is an IBM product.

Table 2. Description of variables

No.	Variable Name	Variable Type	Description	Input/Target
1	Accum_KIPS_Since_Resurfacing	Continuous	Accumulated kips since resurfacing measured in kips	Input
2	Age_2013	Continuous	Years since construction or resurfacing (as calculated from 2013 data)	Input
3	Annual_18_KIPS	Continuous	Annual 18 kips measured in esals	Input
4	Average_Daily_Traffic	Continuous	Average daily traffic as a count per day	Input
5	Average_Daily_Trucks	Continuous	Number of trucks per day	Input
6	DaysTempChange_2013	Continuous	Number of days in 2013 where the maximum temperature was greater than 32F and the minimum temperature was less than or equal to 32F	Input
7	Friction_Value	Continuous	Friction value from 5 to 75	Input
8	IRI_Index	Continuous	International Roughness Index	Input
9	Number_Of_Lanes	Continuous	Number of lanes	Input
10	Pavement_Thickness	Continuous	Pavement thickness in inches	Input
11	Pavement_Width	Continuous	Pavement width	Input
12	Reconstruct_18_KIPS	Continuous	Accumulated kips since construction measured in kips	Input
13	Speed_Limit	Continuous	Speed limit in miles per hour	Input
14	Surface_Type	Continuous	Surface type ranging between 30 and 92	Input
15	PCI_2013	Continuous	Pavement Condition Index in 2013	Input
16	PCI_2014	Continuous	Pavement Condition Index in 2014	Target
17	PCI_2015	Continuous	Pavement Condition Index in 2015	Target
18	Median	Binary	No/Yes with Yes indicating the segment has a median	Input
19	RS_in2013	Binary	No/Yes with Yes indicating the segment was resurfaced in 2013	Input
20	RS_in2014	Binary	No/Yes with Yes indicating the segment was resurfaced in 2014	Input
21	RS_in2015	Binary	No/Yes with Yes indicating the segment was resurfaced in 2015	Input

Table 3. Summary statistics for continuous variables

No.	Variable Name	Portland Cement (n=1251)	Composite (n=1876)	Asphalt Cement (n=462)
			Mean (StdDev)	
1	Accum_KIPS_Since_Resurfacing	105961.99 (865665.45)	1369875.46 (1445019.27)	1346251 (4486387.7)
2	Age_2013	26.54 (17.89)	51.13 (18.9)	34.75 (16.86)
3	Annual_18_KIPS	410225.8 (693750.91)	86587.85 (117240.7)	149744.09 (386827.95)
4	Average_Daily_Traffic	10759.34 (11926.74)	5908.01 (6495.61)	4501.26 (8419.52)
5	Average_Daily_Trucks	1459.27 (2119.88)	521.88 (520.01)	784.43 (1838.57)
6	DaysTempChange_2013	88.95 (32.26)	91.96 (26.64)	86.93 (32.53)
7	Friction_Value	37.62 (23.37)	34.2 (24.78)	40.45 (22.97)
8	IRI_Index	46.57 (22.08)	55.45 (20.42)	59.24 (19.83)
9	Number_Of_Lanes	3.78 (1.21)	2.73 (1.1)	2.47 (1.02)
10	Pavement_Thickness	10 (1.49)	13.88 (2.83)	11.97 (4.34)
11	Pavement_Width	26.8 (8.62)	27.61 (8.08)	24.79 (5.06)
12	Reconstruct_18_KIPS	9627721.86 (16333908.38)	4579669.22 (4339765.43)	5489661.83 (14416938.21)
13	Speed_Limit	54.54 (12.28)	51.34 (9.36)	54.59 (7.55)
14	Surface_Type	73.75 (2.14)	67.1 (6.53)	63.62 (6.07)
15	PCI_2013	61.93 (16.42)	64.91 (17.01)	64.92 (18.42)
16	PCI_2014	64.82 (18.16)	66.9 (19.38)	67.11 (19.41)
17	PCI_2015	66.32 (22.37)	65.3 (21.98)	67.08 (20.82)

Table 4. Summary statistics for binary variables

		Portland Cement (n=1251)		Composite (n=1876)		Asphalt Cement (n=462)	
No.	**Variable Name**	**Yes (%)**	**No (%)**	**Yes (%)**	**No (%)**	**Yes (%)**	**No (%)**
18	Median	64.03	35.97	21.86	78.14	16.67	83.33
19	RS_in2013	0	100	3.36	96.64	4.55	95.45
20	RS_in2014	1.76	98.24	3.2	96.8	1.73	98.27
21	RS_in2015	0.8	99.2	1.65	98.35	3.03	96.97

2. MODELING PCI FOR PORTLAND CEMENT PAVEMENT TYPE

Data Quality

IBM Watson Analytics provides a score between 1 and 100 as a measure of the overall quality of the data set being used in analysis. It also flags variables relative to their quality scores. The quality scores are, in general, determined by the percent of missing values, extent to which the values of variables vary, and several other factors. Figure 1 shows the overall data quality score as determined by Watson Analytics for this data set (portland cement). A score of 59 is considered to be medium quality.

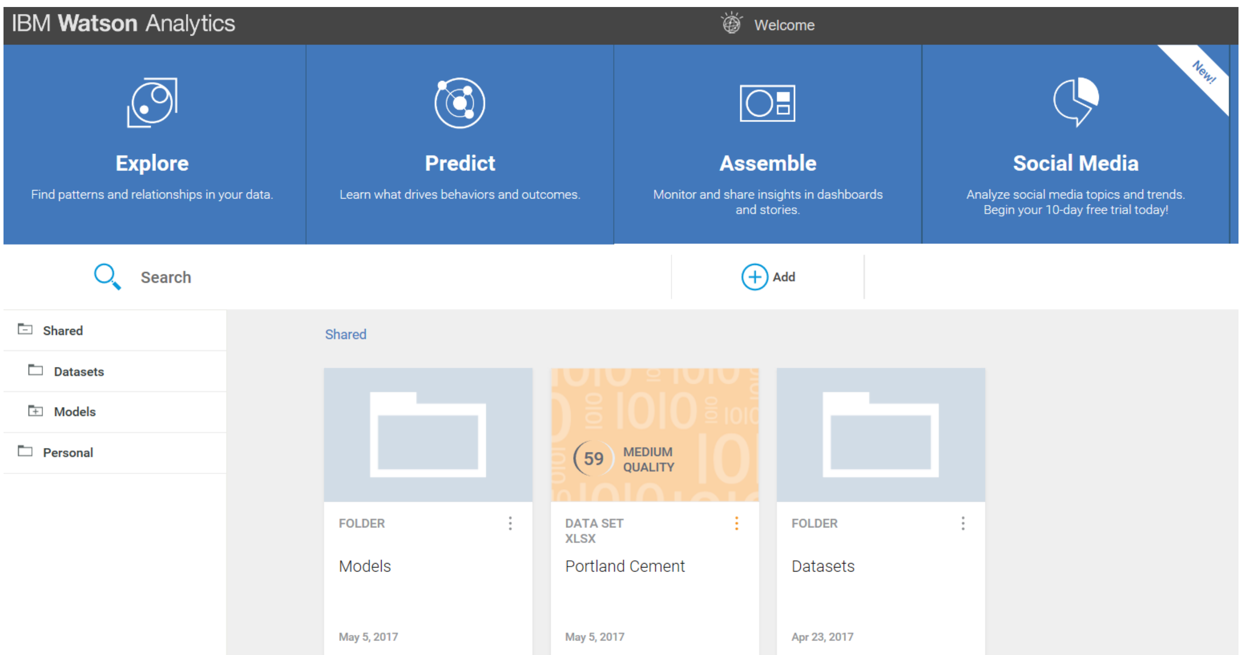

Figure 1. Screenshot for Watson Analytics page of overall data set for portland cement

Figures 2 and 3 identify the high-quality and low-quality variables, respectively. Variables RS_2015 and Accum_KIPS_Since_Resurfacing are shown as having the lowest data quality, while variable Number_of_Lanes has the highest data quality score, 93 (Figure 2).

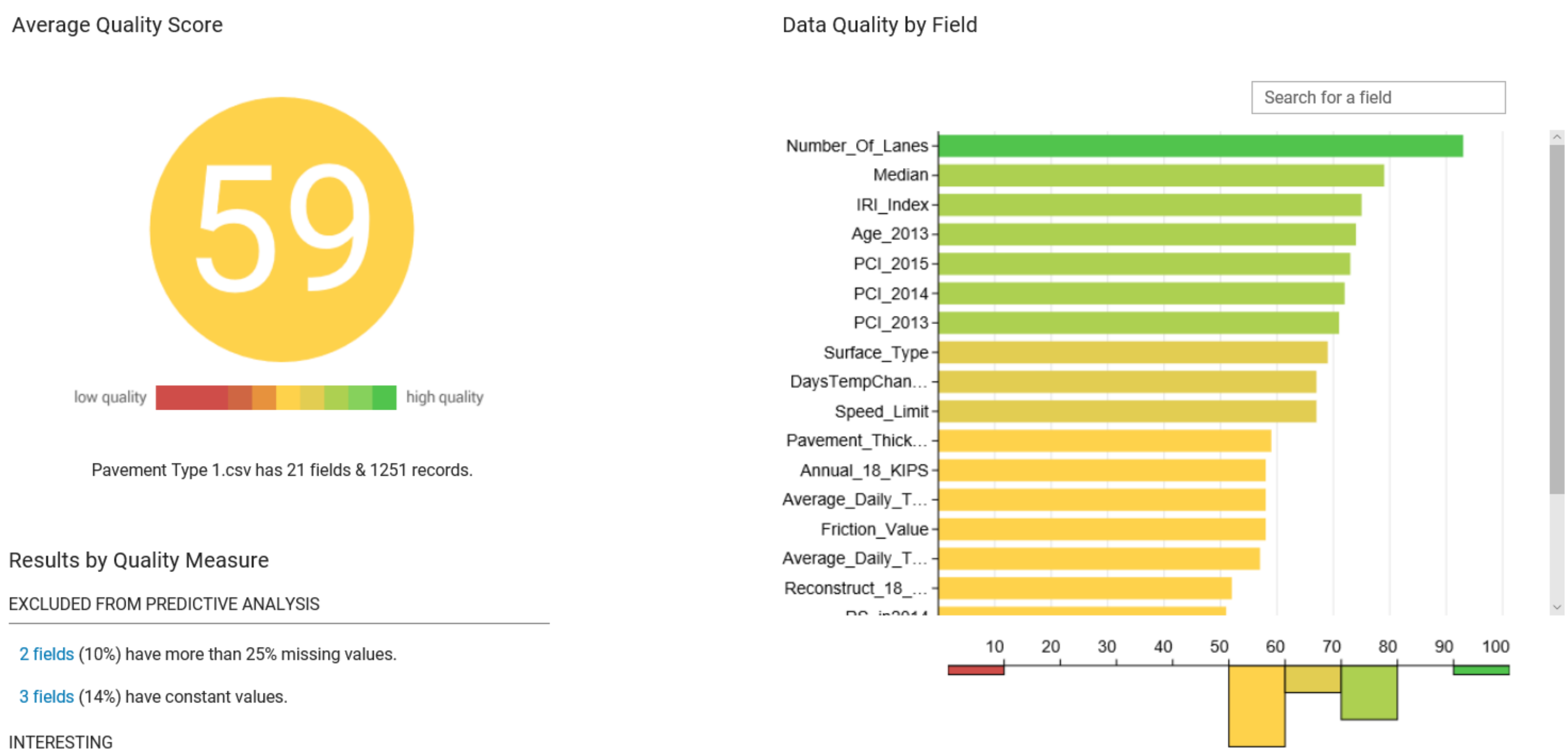

Figure 2. Screenshot for high-quality variables for portland cement

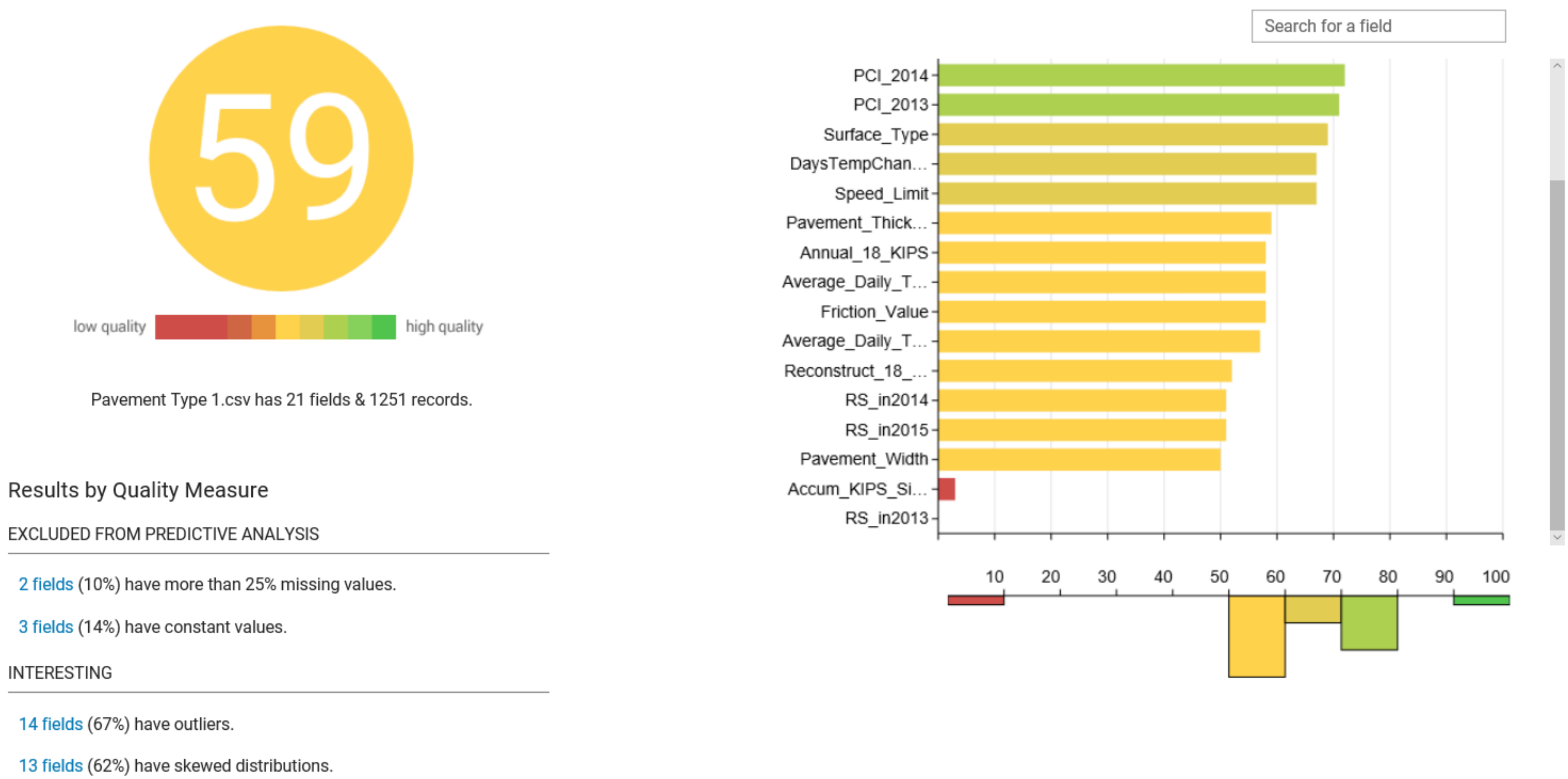

Figure 3. Screenshot for low-quality variables for portland cement

Predicting PCI_2014 for Portland Cement Pavement Type

The following steps show the process for creating a model to predict PCI_2014 using Watson Analytics.

Step1. Click on the data set and then click on the "Predict" icon.

Figures 4 and 5 demonstrate how to start a new analysis by choosing a data set and then using that to predict PCI_2014.

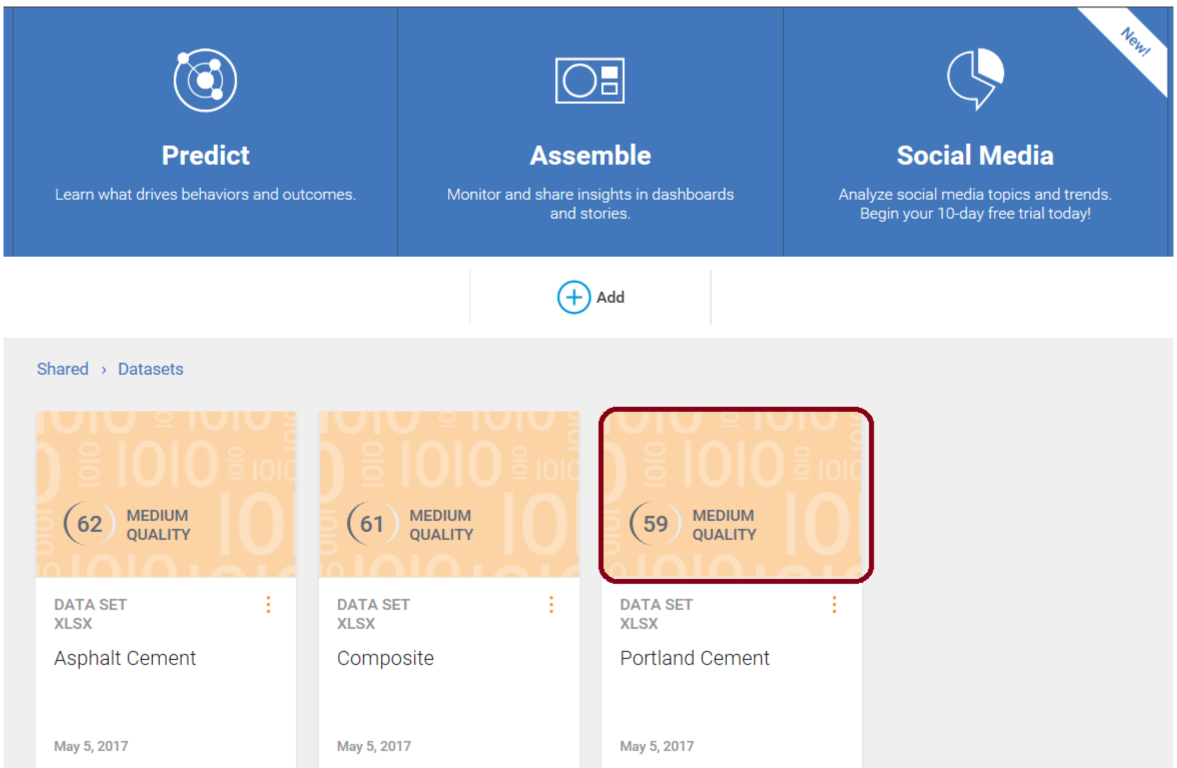

Figure 4. Screenshot highlighting selection of portland cement data set

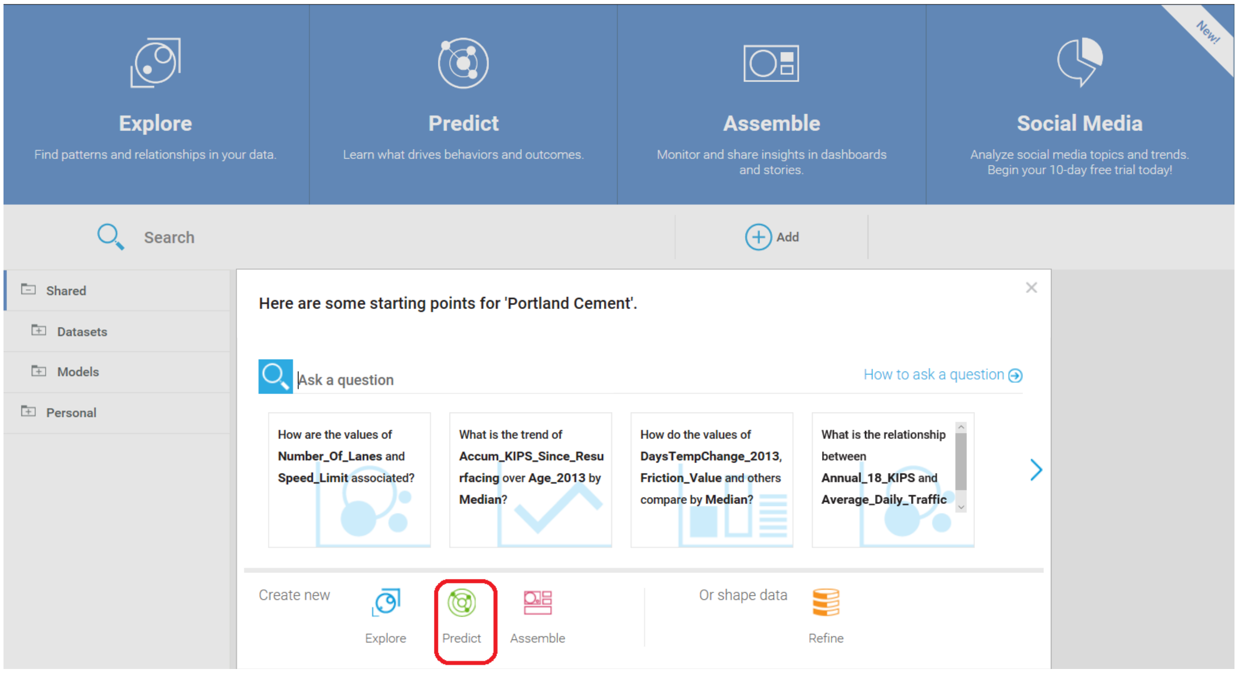

Figure 5. Screenshot highlighting predict icon

Step 2. Enter a workbook name and then select "Edit this workbook's field properties" (Figure 6) to select variables that would be used as input and target.

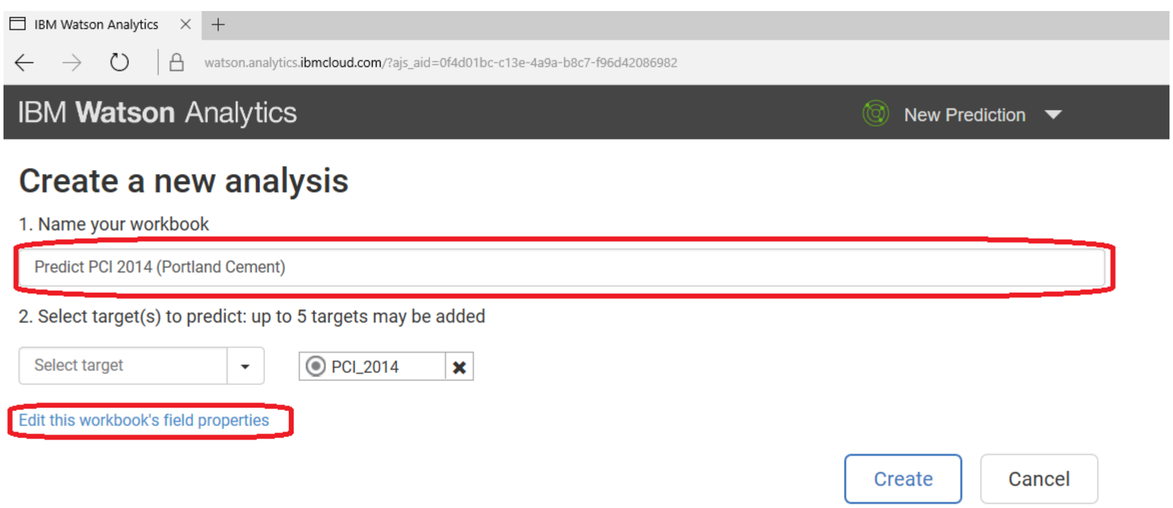

Figure 6. Screenshot highlighting workbook name and edit function

Step 3. Select PCI_2014 as the target and include 18 variables as inputs by excluding PCI_2015 and RS_2015 from the model, since these two variables are not relevant when predicting 2014 PCI. Then, select "Continue" (Figure 7).

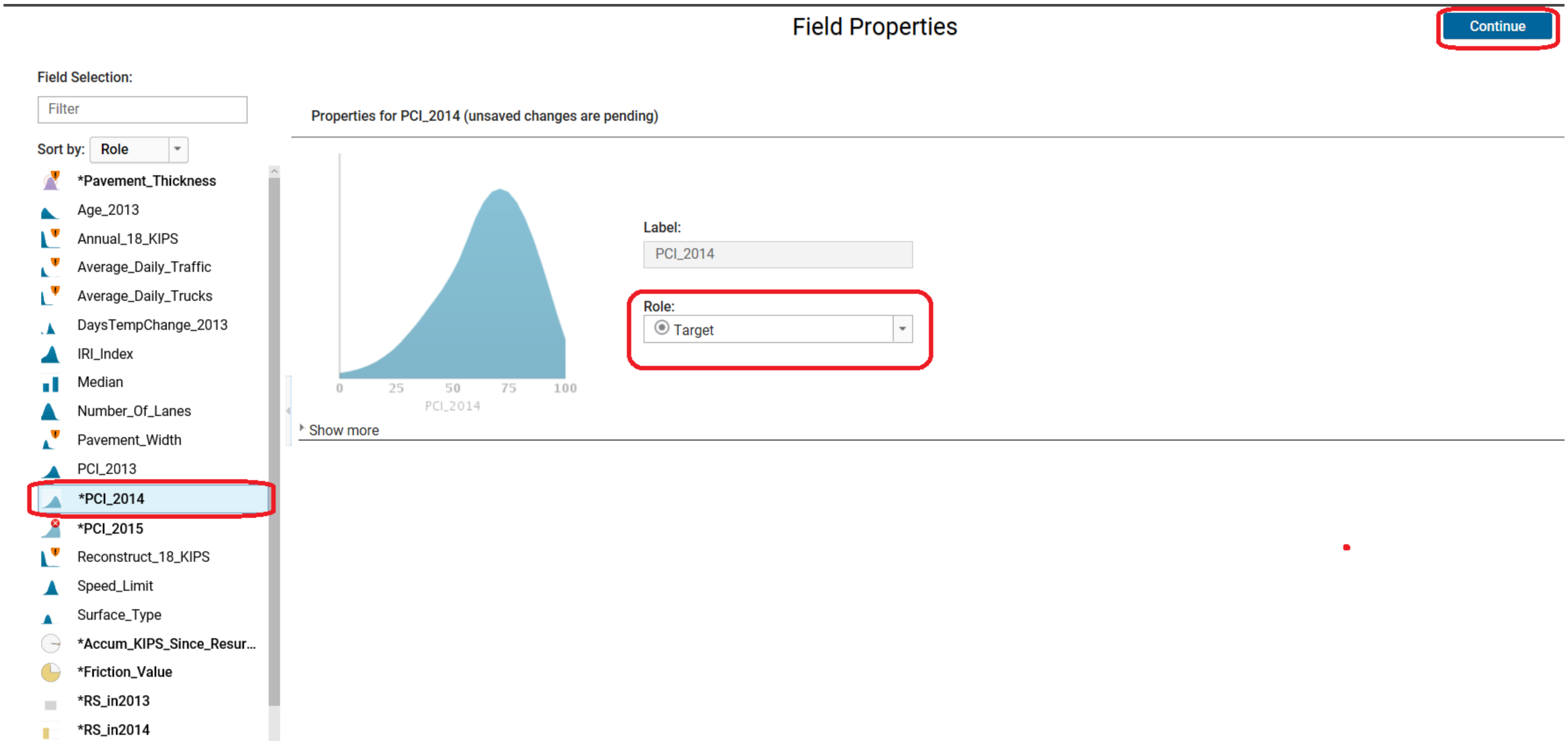

Figure 7. Screenshot highlighting selection of PCI_2014 as target

Step 4. The screen shown in Figure 6 will appear. Then, select "Create" and wait until the new prediction workbook is created.

Figure 8 shows the IBM Watson Analytics page that appears as the new workbook is being created.

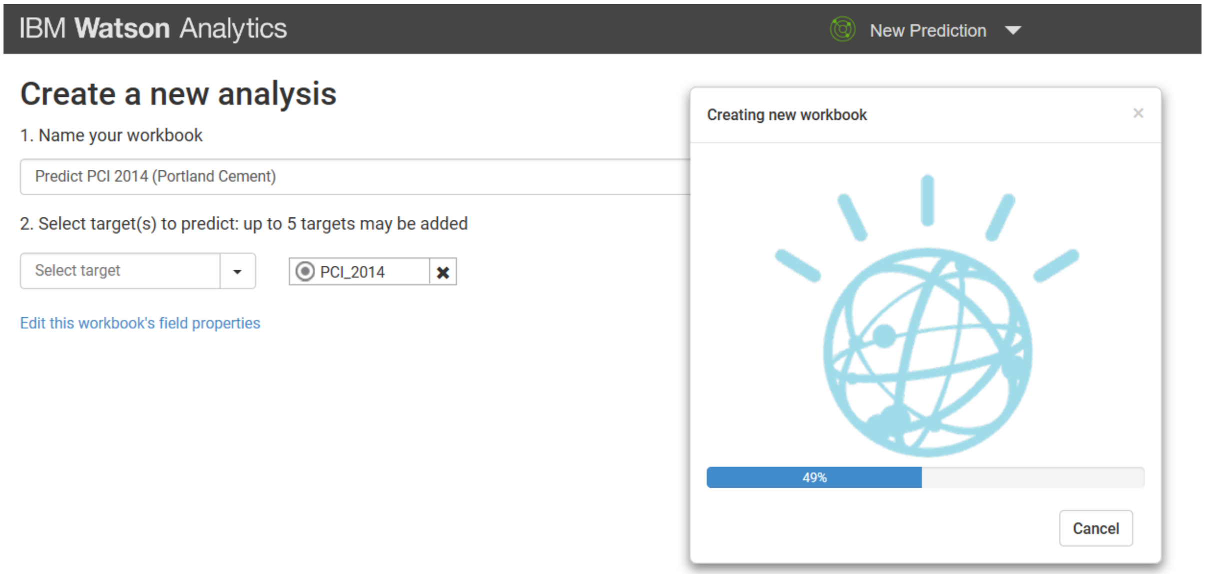

Figure 8. Screenshot of page after selecting "Create" to make a new analysis

Step 5. When a new workbook is created, select "View" on the Top Field Associations section to see fields with strong associations and correlations (Figures 9 and 10).

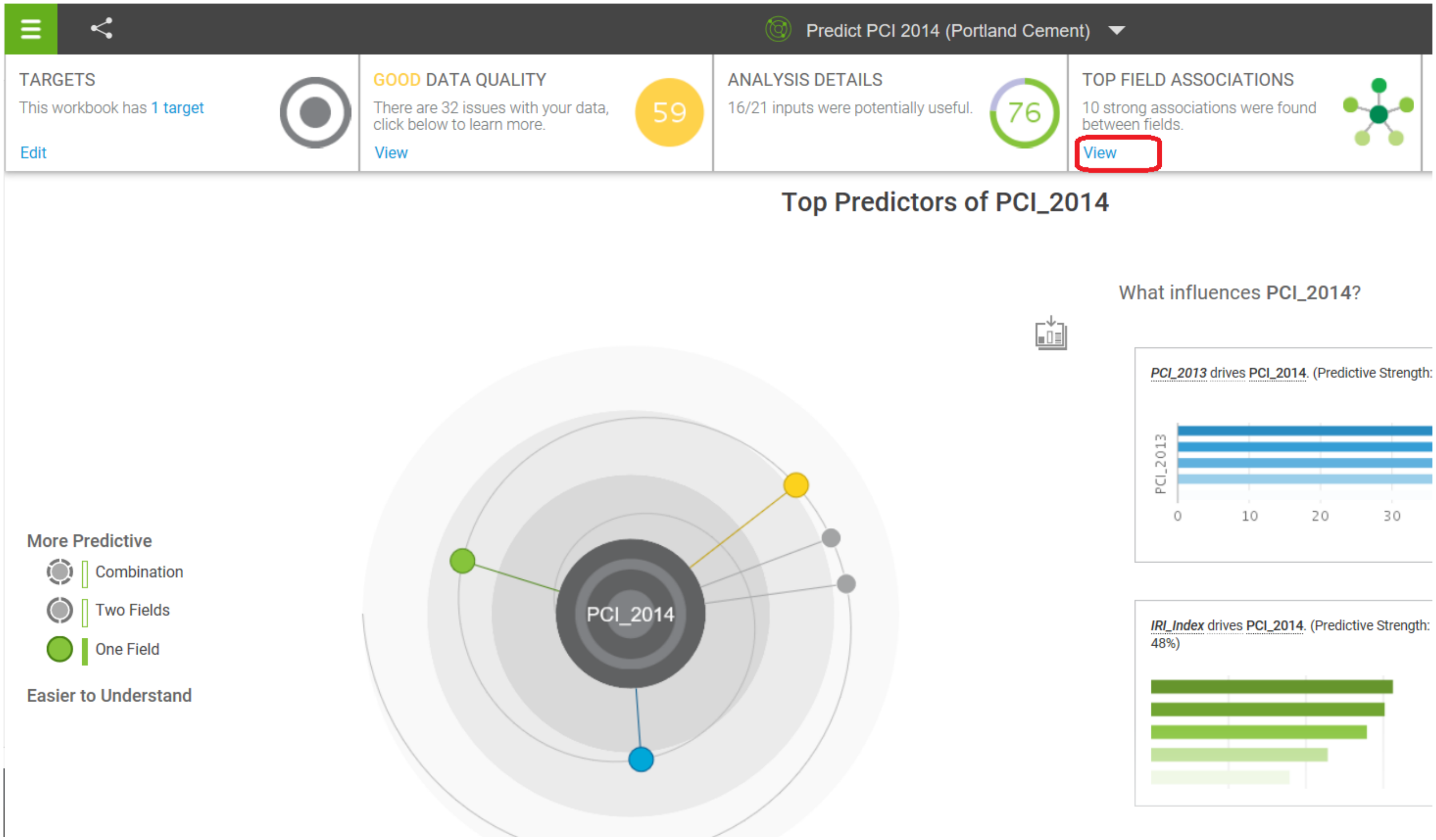

Figure 9. Screenshot highlighting how to view 10 variables with strong field associations for predicting PCI_2014

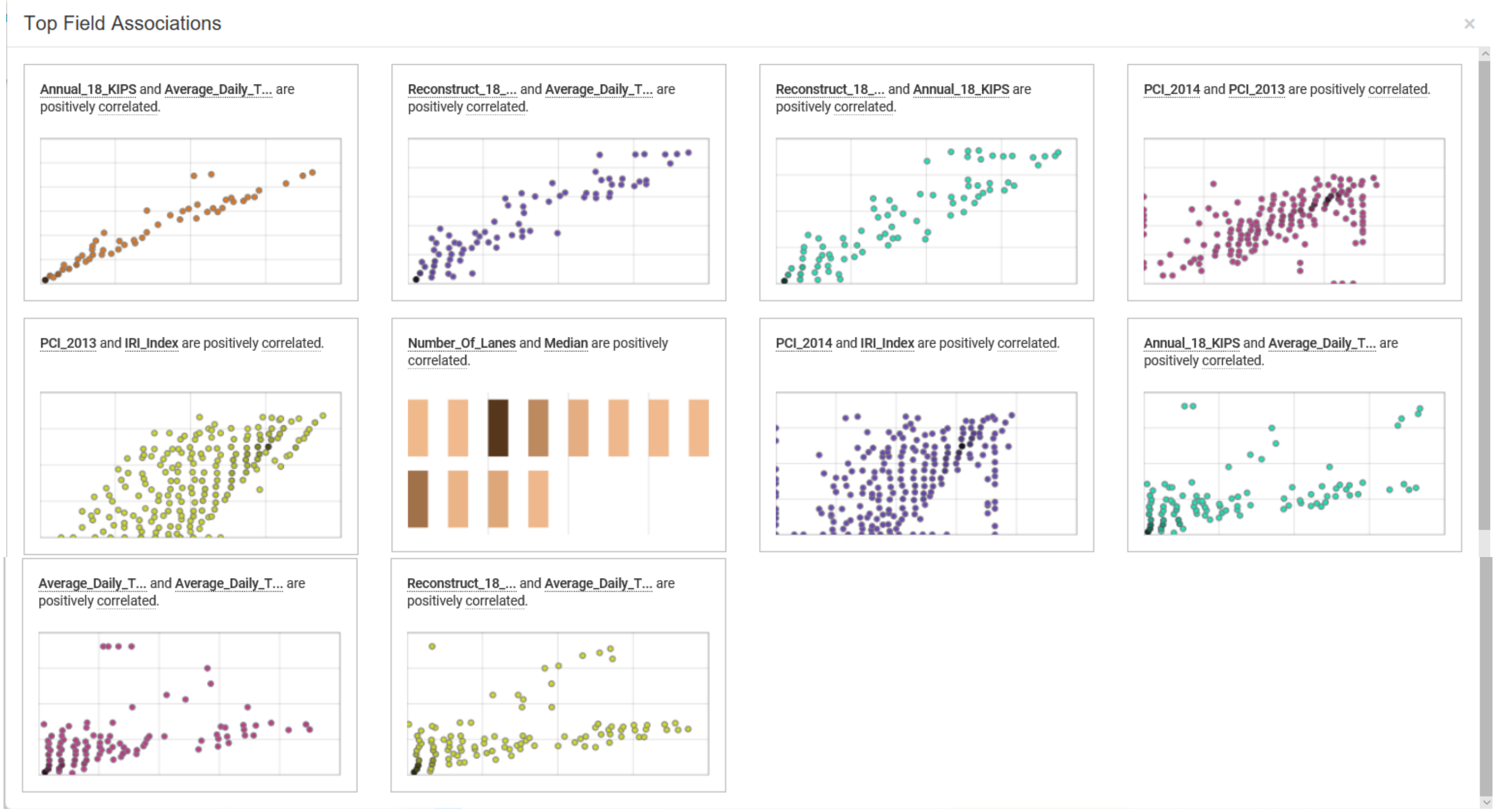

Figure 10. Screenshot of the top 10 strong field associations

Step 6. Select "Two Fields" and "Combination" in the More Predictive section (Figure 11). This step includes combinations of variables that are strong predictors of PCI_2014.

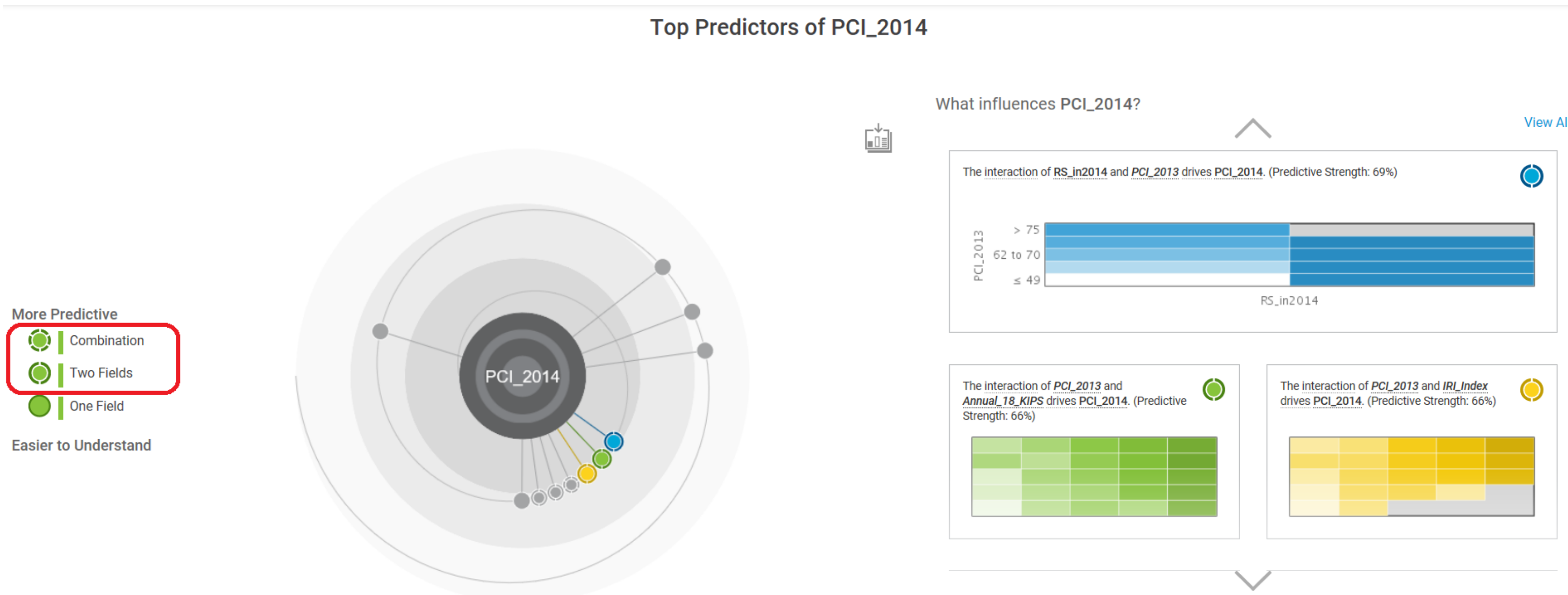

Figure 11. Screenshot highlighting additional options for predictors

Results

IBM Watson Analytics uses different machine learning (ML) techniques for predictions. One variable, or a combination of variables, can be a strong predictor of the target variable. Figure 12 shows the six variables sized in proportional to their importance in predicting PCI_2014. These variables are PCI_2013, Annual_18_KIPS, Speed_Limit, IRI_Index, Pavement_Width, and Age_2013.

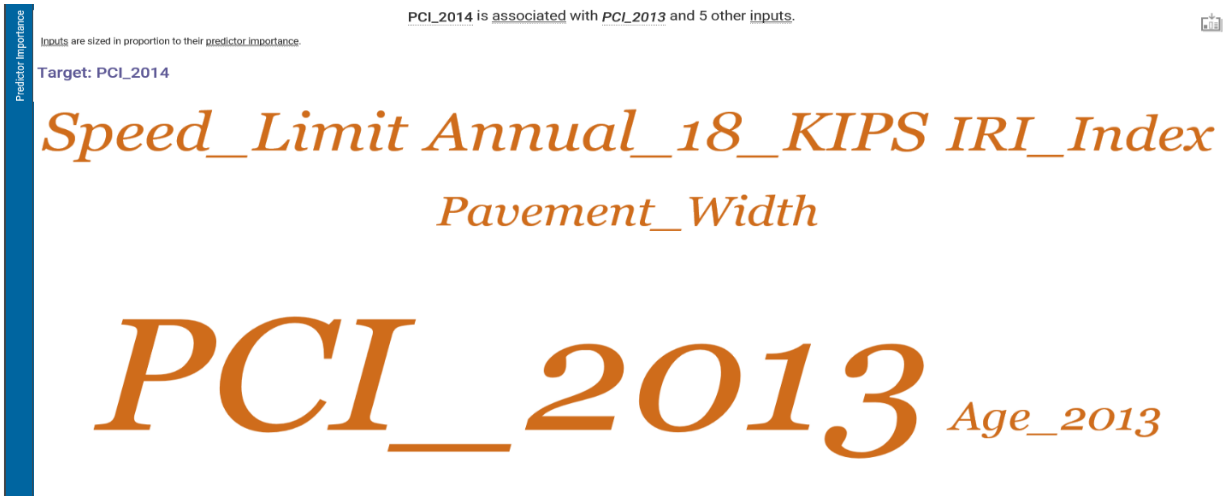

Figure 12. Screenshot of word cloud showing six variables sized in proportion to their importance in predicting PCI_2014

Figures 13 through 17 show the predictive strengths of five of the six variables when looking at "One Field" outcomes. The predictive strength of the sixth variable (Pavement_Width) can be assessed when looking at the "Two Fields" and "Combination" results. Note that PCI_2013 shows a predictive strength of 63.1%, IRI_Index shows a predictive strength of 47.6%, Age_2013 shows a predictive strength of 30.3%, Annual_18_KIPS shows a predictive strength of 27.9%, and Speed_Limit shows a predictive strength of 26.1%. Further, the interaction between Age_2013 and Annual_18_KIPs shows a predictive strength of 42.2%, and the interaction between Average_Daily_Traffic and Speed_Limit yields a predictive strength of 32.8%.

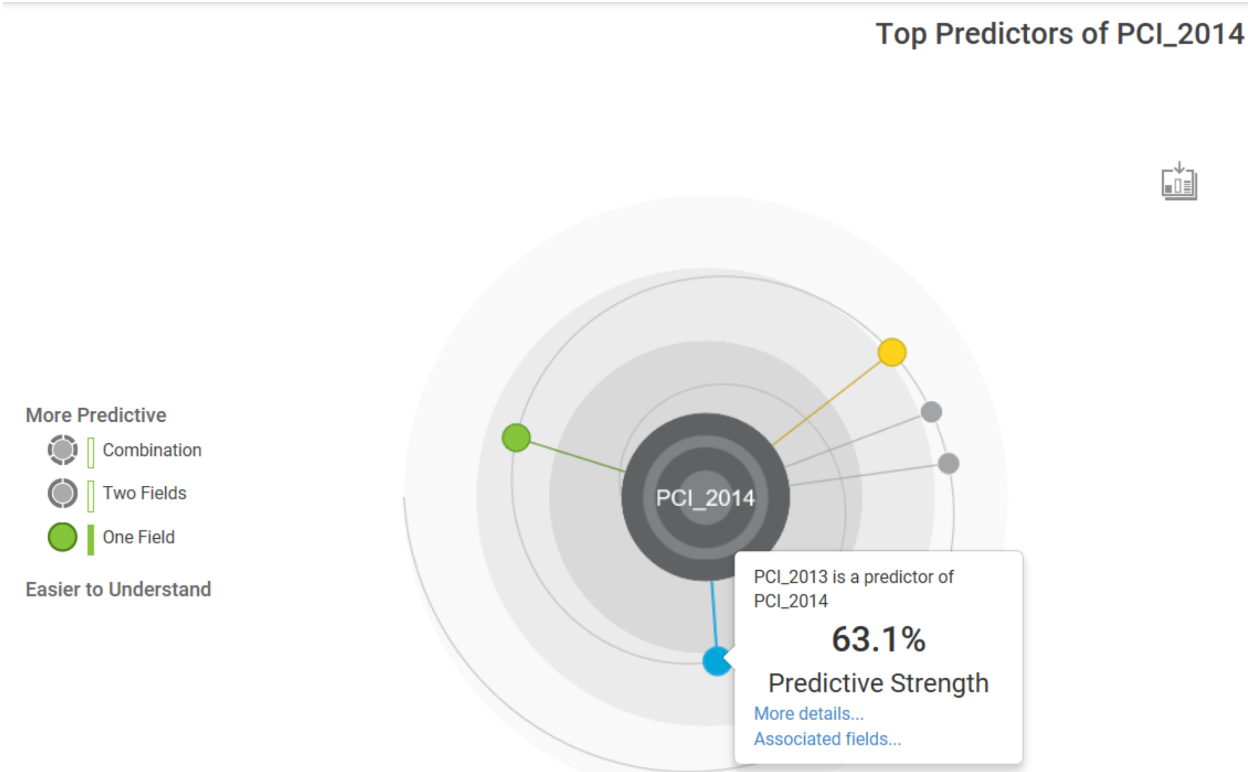

Figure 13. Screenshot of predictive strength of the PCI_2013 variable for PCI_2014

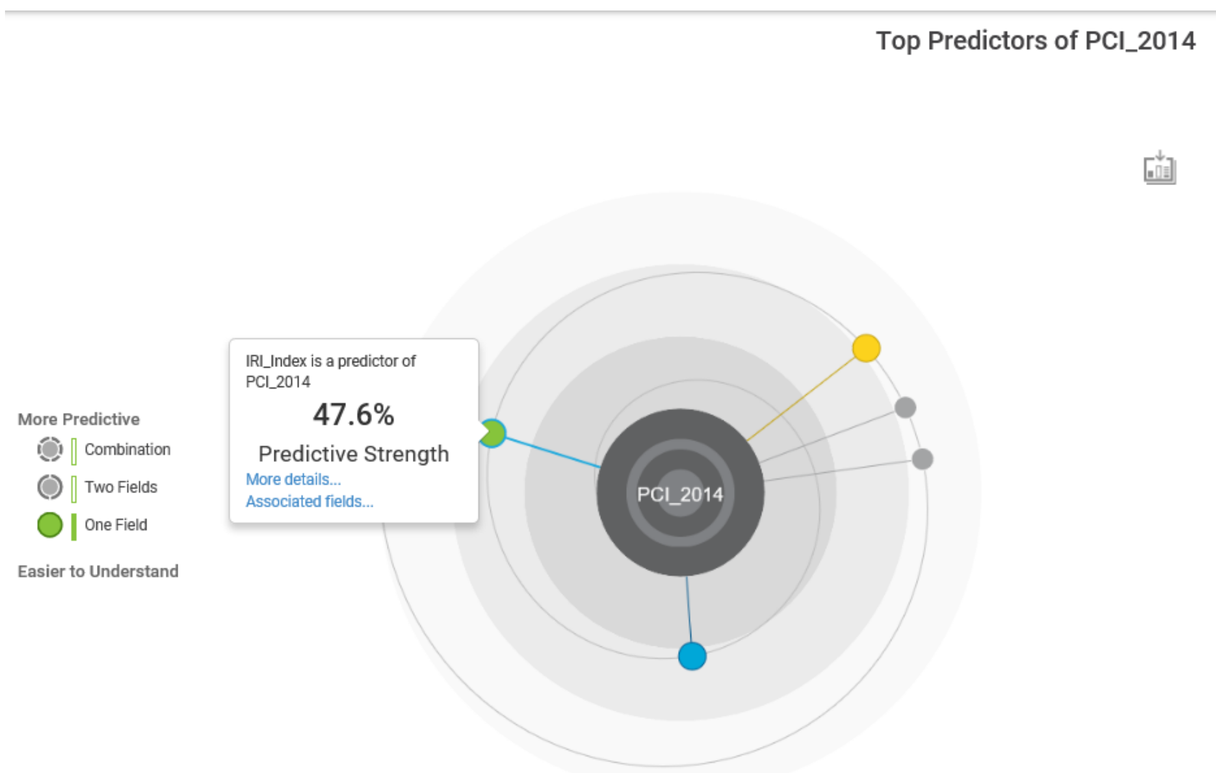

Figure 14. Screenshot of predictive strength of the IRI_Index variable for PCI_2014

Figure 15. Screenshot of predictive strength of the Speed_Limit variable for PCI_2014

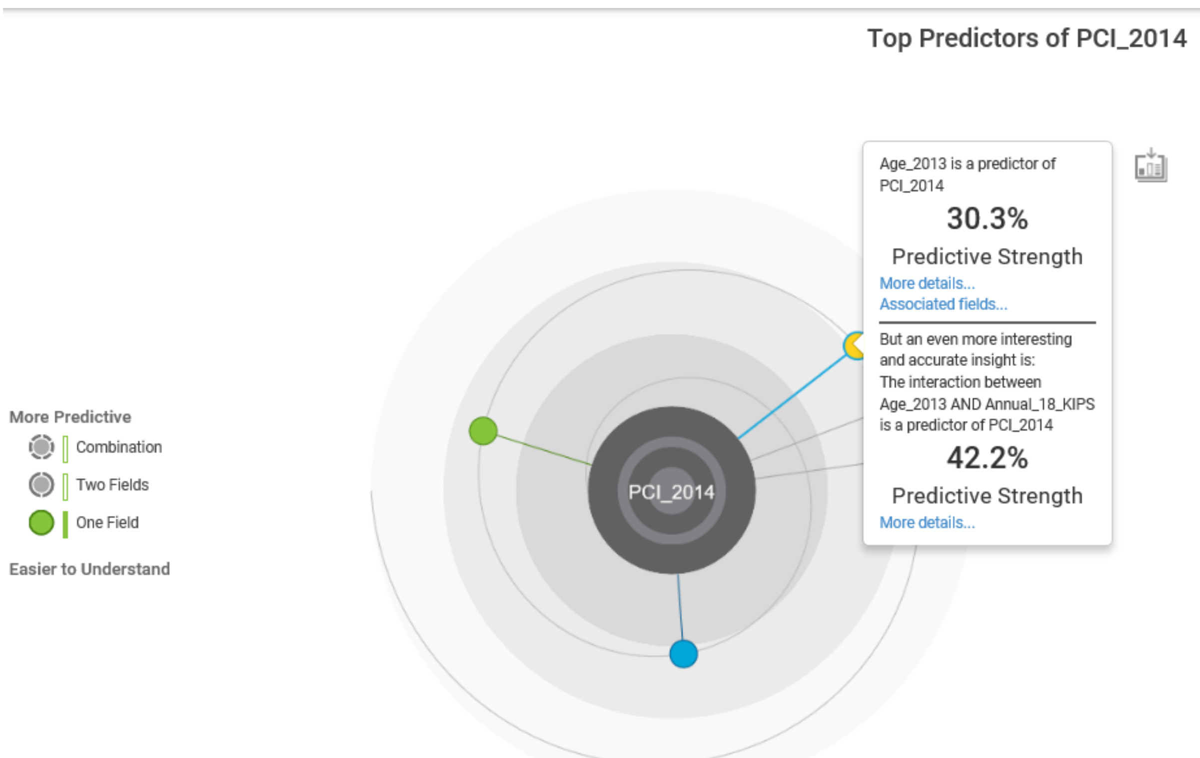

Figure 16. Screenshot of predictive strength of the Age_2013 variable for PCI_2014

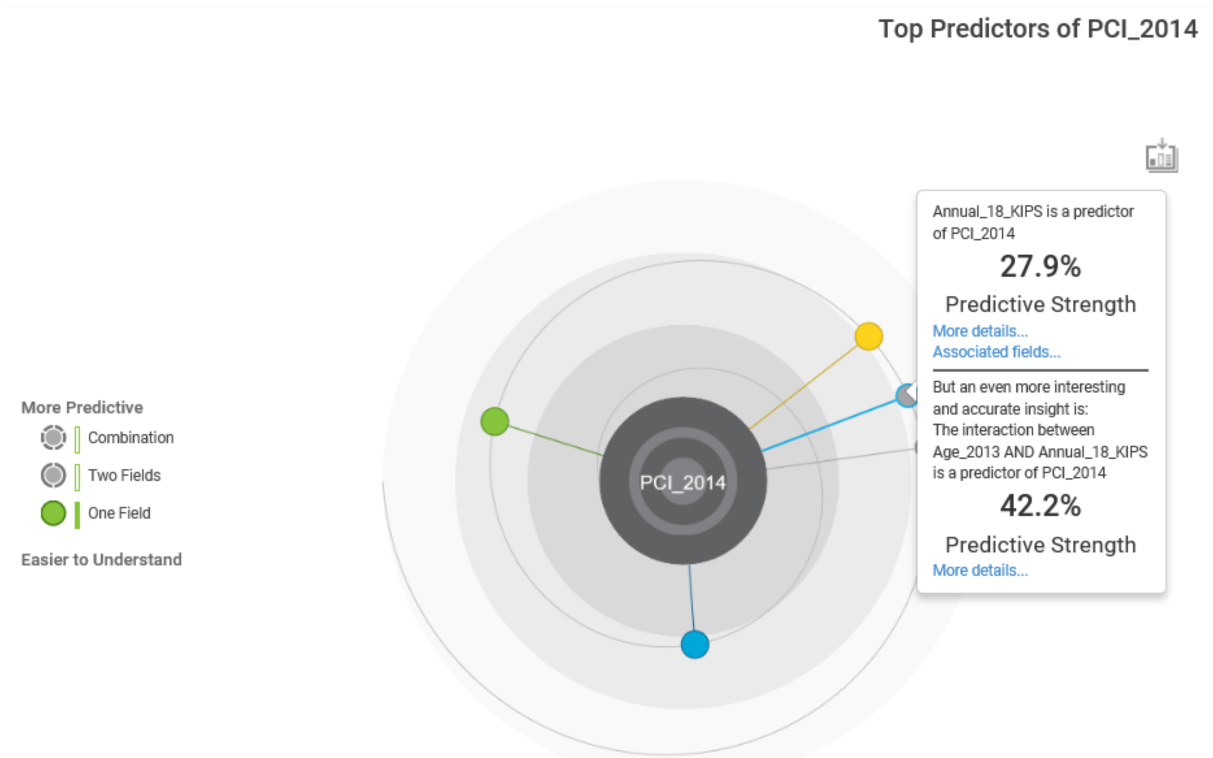

Figure 17. Screenshot of predictive strength of the Annual_18_KIPS variable for PCI_2014

Figure 18 shows a decision tree produced by Watson Analytics that depicts the associative rules for the six predictors and the outcome variable PCI_2014.

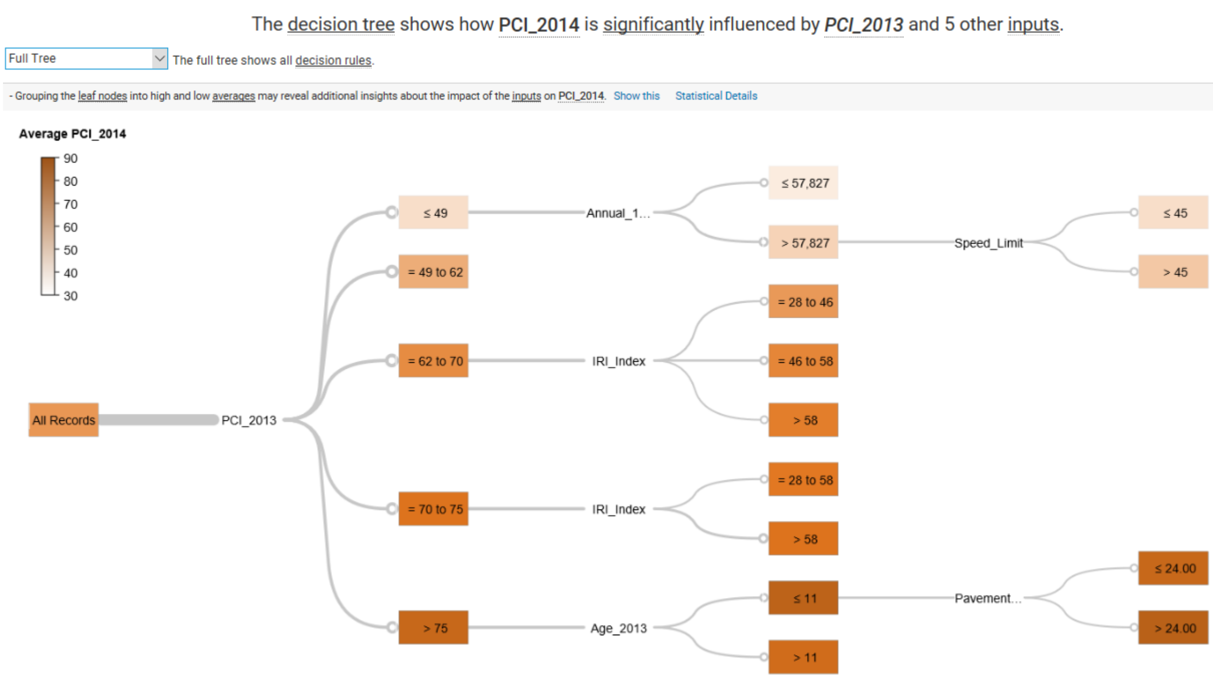

Figure 18. Screenshot of decision tree showing associative rules for six predictors

3. MODELING PCI FOR COMPOSITE PAVEMENT TYPE

The previous chapter, which predicts 2014 PCI for the portland cement pavement type, provides substantial details of the procedure for predicting 2014 PCI using Watson Analytics. For the other pavement types, described in the Chapters 3 through 5, most of the details are omitted and only the final results are shown.

Data Quality

Watson Analytics rated the overall data quality for the composite pavement type as medium, with a score of 61, which is slightly higher than the score of 59 for portland cement pavement. For individual variables, Age_2013 has the best data quality score, 83, while the rest of the variables have data quality scores ranging from 50 to 79.

Predicting PCI for Composite Pavement Type

Figure 19 shows a word cloud for the eight significant input variables. The eight variables are PCI_2013, IRI_Index, Pavement_Width, Median (present/ absent), Number_Of_Lanes, Reconstruct_18_KIPS, Accum_KIPS_Since_Resurfacing, and Age_2013.

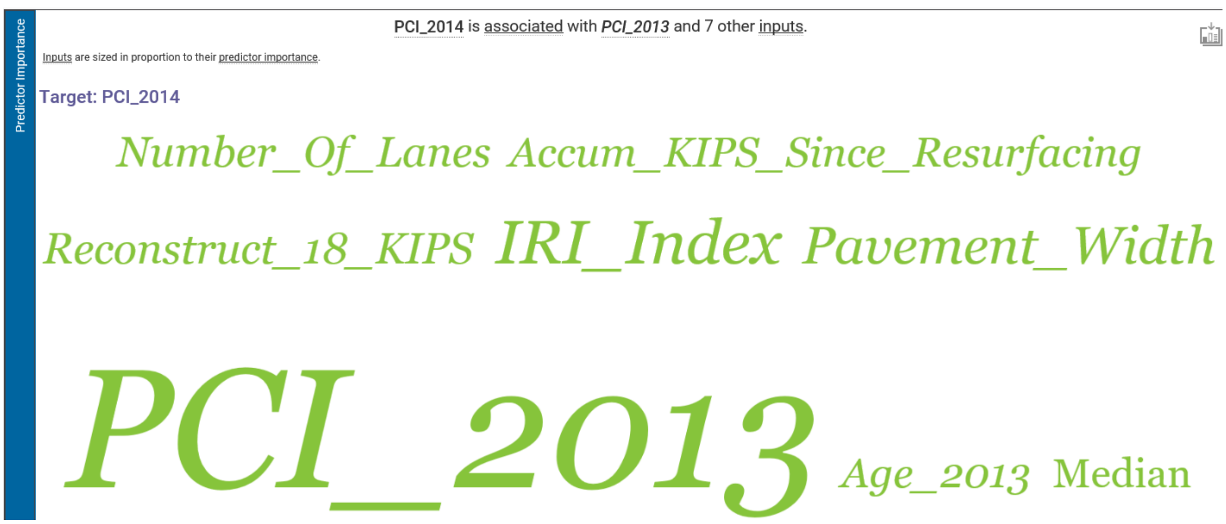

Figure 19. Screenshot of word cloud showing the eight significant variables to predict PCI_2014

Figure 20 shows the decision tree that depicts the extent to which the top input variables influence and predict PCI_2014.

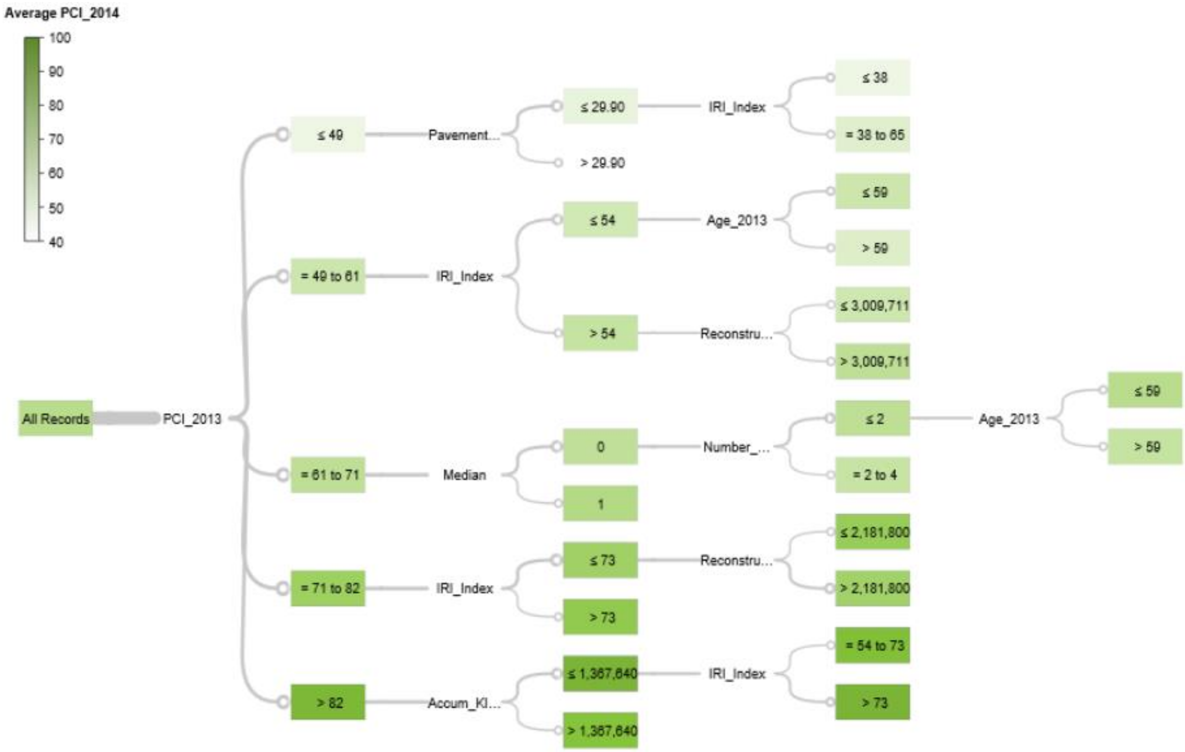

Figure 20. Screenshot of decision tree that shows the extent to which the top input variables predict PCI_2014

4. MODELING PCI FOR ASPHALT CEMENT PAVEMENT TYPE

Data Quality

IBM Watson Analytics classified the overall data quality for the asphalt cement pavement type as medium, with a score of 62. This is slightly better than the data quality scores for the portland cement (59) and composite (61) pavement types. The Age_2013 variable has the best data quality score, 93, while other variables have data quality scores ranging from 37 to 86.

Predicting PCI_2014 for Asphalt Cement Pavement Type

The analysis showed that only one input variable, PCI_2013, is significant in predicting PCI_2014. Figure 21 shows a decision tree outlining the rules for predicting PCI_2014 based upon the values of PCI_2013.

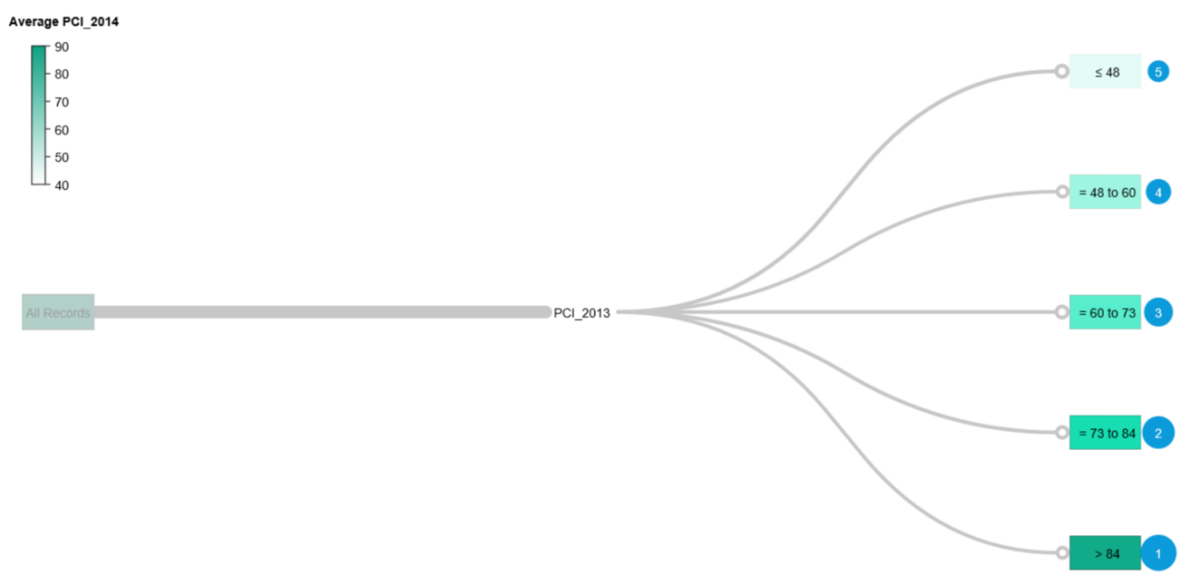

Figure 21. Screenshot of full decision tree showing the rules for predicting PCI_2014

Apart from PCI_2013, which is a strong predictor of PCI_2014, the analysis shows that there are four combinations of variables that can be used for this prediction (Figures 22 through 25). These four combinations include PCI_2013 and RS_In2014 (74.7% predictive strength), PCI_2013 and Annual_18_KIPS (71.2% predictive strength), PCI_2013 and Average_Daily_Traffic (70.3% predictive strength), and PCI_2013 and Average_Daily_Trucks (70.1% predictive strength).

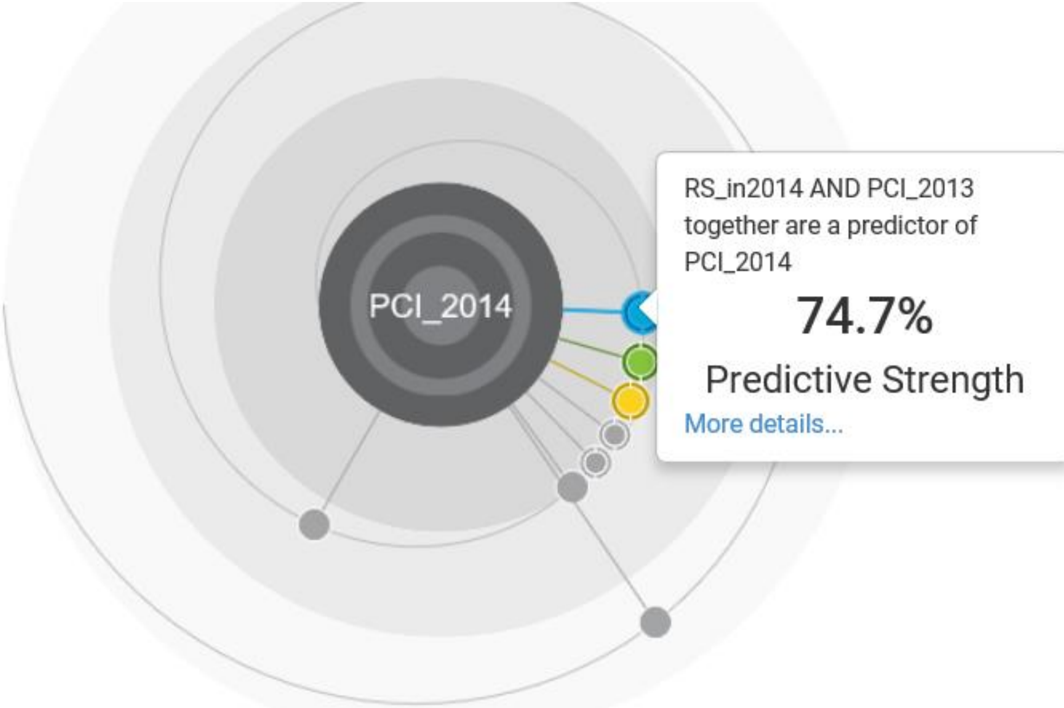

Figure 22. Screenshot of the RS_in2014 and PCI_2013 combination of input variables for predicting PCI_2014

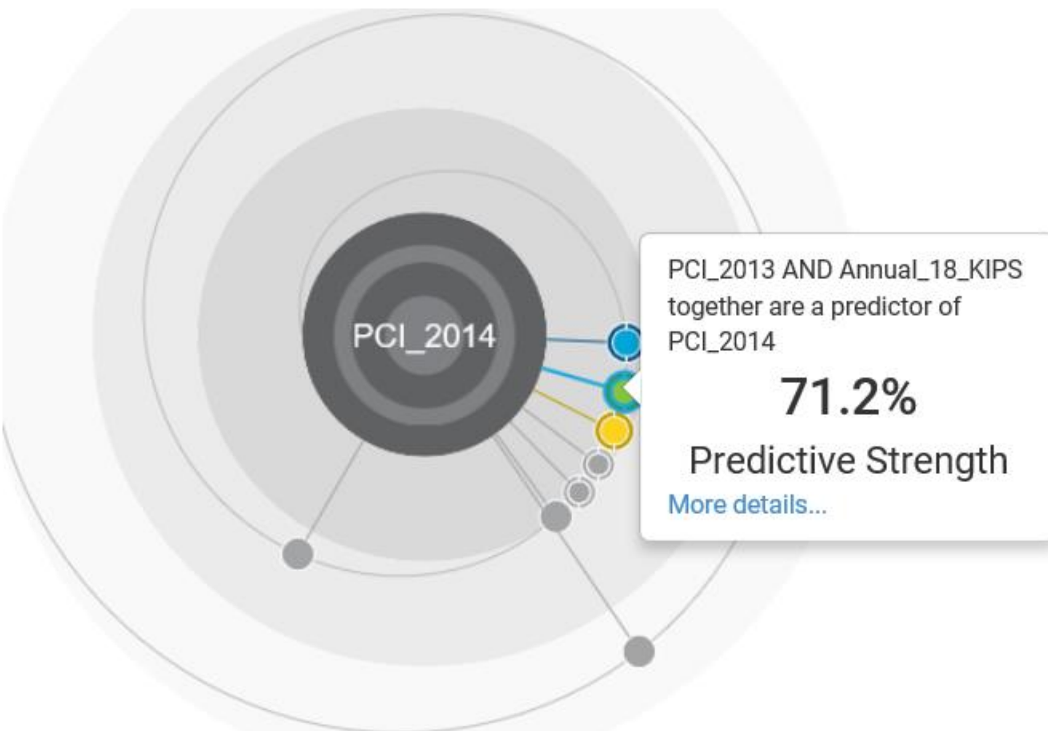

Figure 23. Screenshot of the PCI_2013 and Annual_18_KIPS combination of input variables for predicting PCI_2014

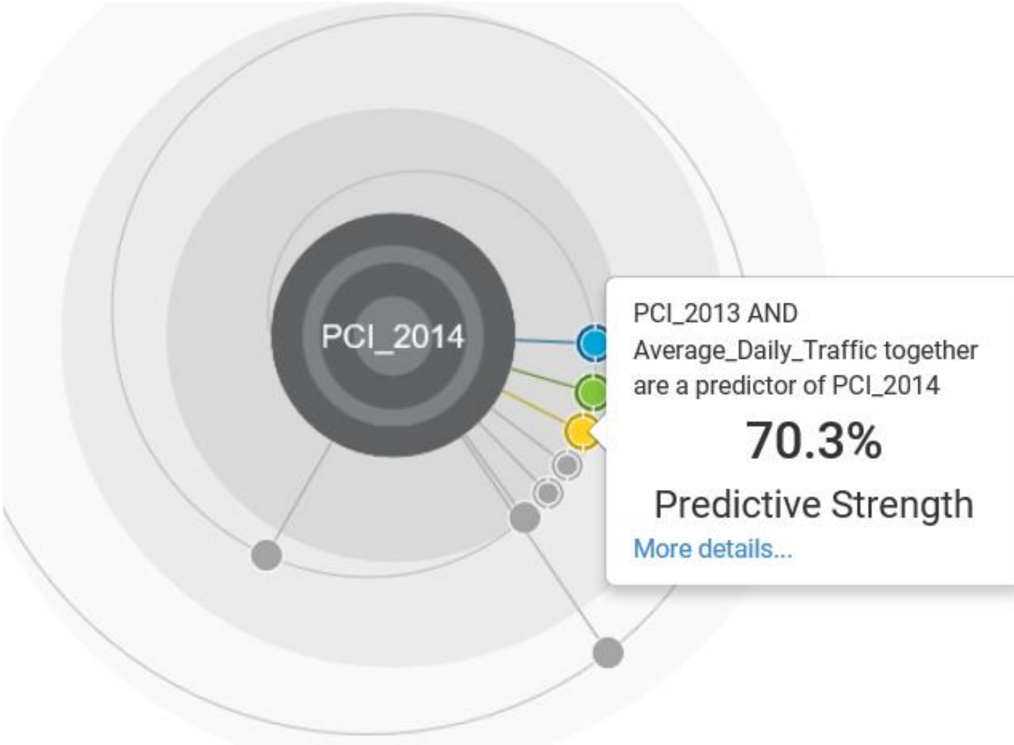

Figure 24. Screenshot of the PCI_2013 and Average_Daily_Traffic combination of input variables for predicting PCI_2014

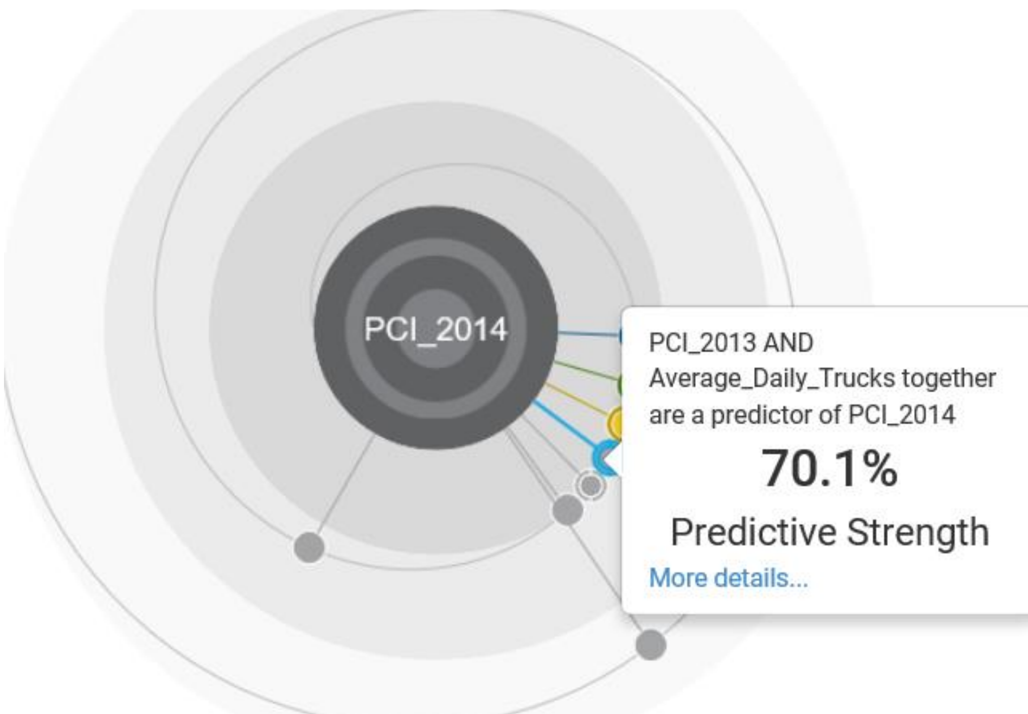

Figure 25. Screenshot of PCI_2013 and Average_Daily_Trucks combination of input variables for predicting PCI_2014

5. PREDICTING TWO YEARS AHEAD (PCI_2015)

Turning our attention to predicting PCI two years ahead (PCI_2015), we selected the PCI_2015 variable as our target and included 19 variables as inputs while excluding the PCI_2014 variable from the model.

Portland Cement

Figure 26 shows a word cloud of the four significant variables for predicting PCI_2015. A combination of PCI_2013, IRI_Index, Average_Daily_Traffic, and Age_2013 are strong predictors of PCI_2015, with a 44% predictive strength.

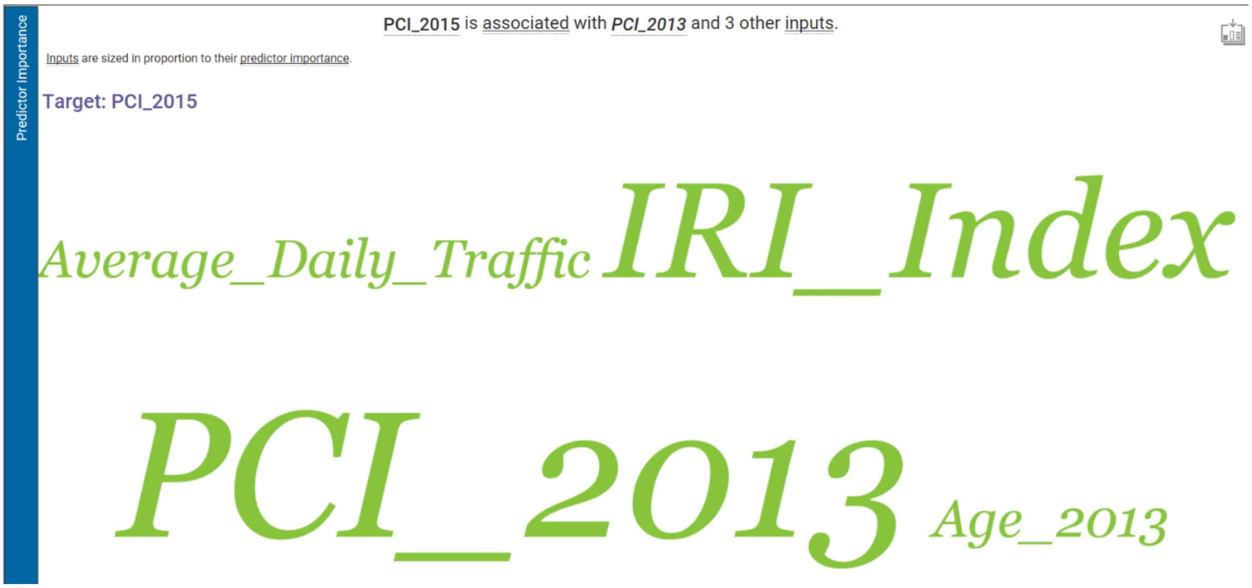

Figure 26. Screenshot of word cloud showing four variables sized in proportion to their importance in predicting PCI_2015

Figure 27 shows a breakdown of decision tree rules for predicting PCI in 2015 for the portland cement pavement type. The first rule states that to predict two years ahead and to achieve a higher PCI in 2015, PCI_2013 should be greater than 70, Average_Daily_Traffic must be greater than 6,400 vehicles, and IRI_index must be greater than 59.

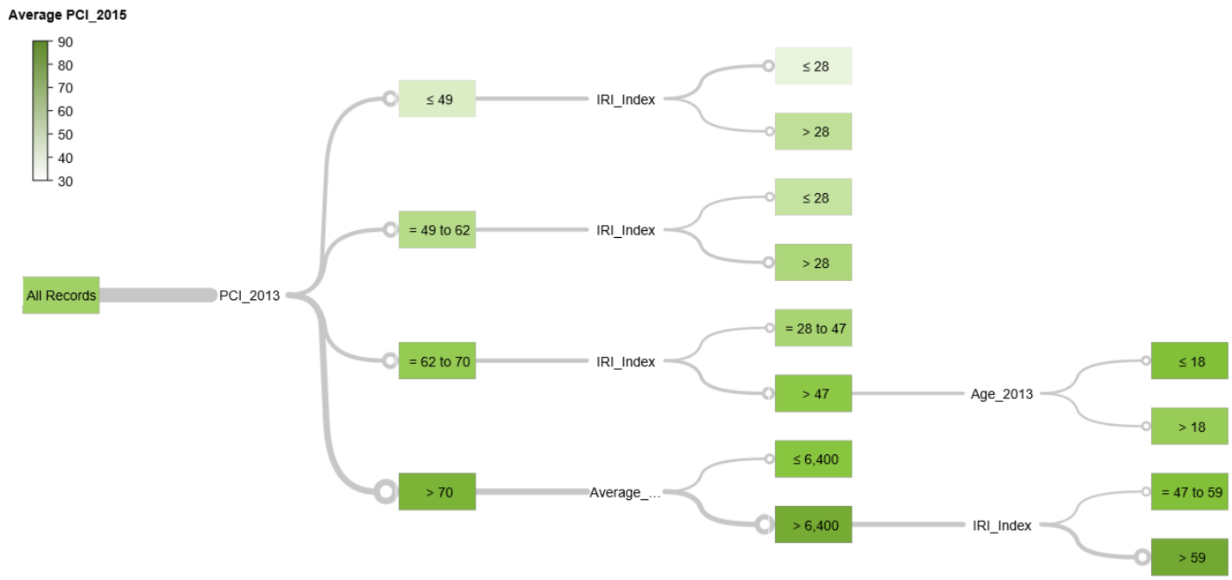

Figure 27. Screenshot of decision tree rules for predicting PCI_2015 for portland cement

Composite

Figure 28 shows a word cloud of the five significant input variables for predicting PCI_2015.

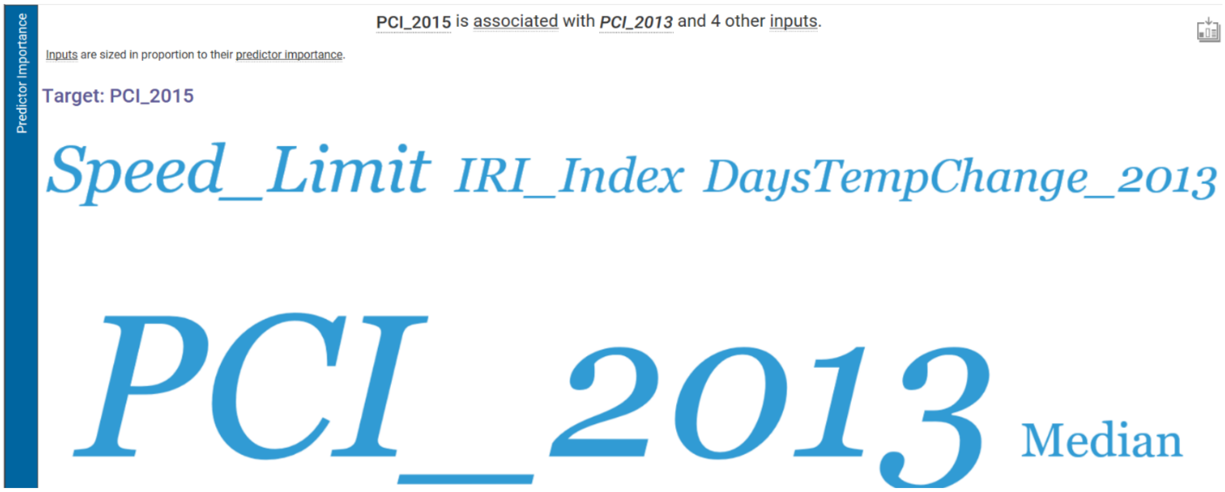

Figure 28. Screenshot of word cloud showing five variables sized in proportion to their importance in predicting PCI_2015 for composite

Figure 29 shows a breakdown of decision tree rules for predicting PCI in 2015 for the composite pavement type. The first rule states that to predict two years ahead and to achieve a high PCI in 2015, PCI_2013 should be greater than 82, Median must be 1 (Yes), and IRI_Index must be greater than 65.

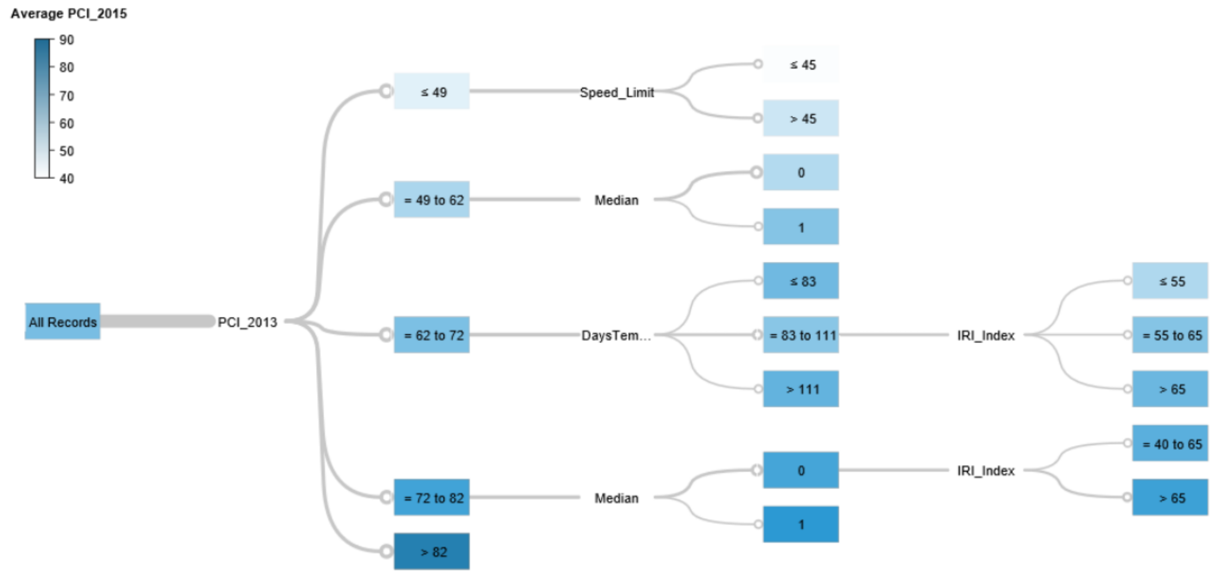

Figure 29. Screenshot of five decision tree rules for predicting PCI_2015 for composite

Asphalt Cement

Figure 30 shows the combined effects of PCI_2013 and RS_In2015 on PCI_2015, which together have a 62% predictive strength.

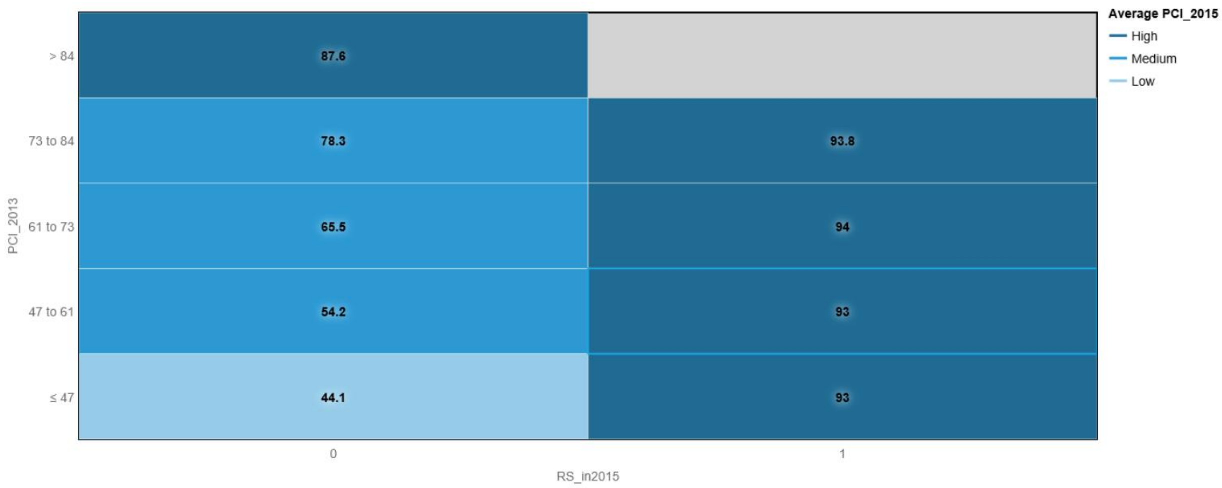

Figure 30. Screenshot of the interaction between PCI_2013 and RS_In2015 as predictors of PCI_2015 for asphalt cement

In this figure, each cell represents the average PCI_2015 for a combination of PCI_2013 and RS_In2015. For example, a “high” value of PCI_2015 is achieved when RS_In2015 is equal to zero and PCI_2013 is greater than 84.

Furthermore, four different combinations of variables also predicted PCI_2015. These four combinations include PCI_2013 and Annual_18_KIPS (53.6% predictive strength), PCI_2013 and Average_Daily_Traffic (52.6% predictive strength), PCI_2013 and Friction_Value (52.3% predictive strength), and PCI_2013 and Average_Daily_Trucks (51.8% predictive strength).

6. SUMMARY

Table 5 summarizes the results of predictive modeling for PCI_2014 as well as PCI_2015. It shows the top input variables for each of the three pavement types (portland cement, composite, and asphalt cement).

Table 5. Key predictors of PCI

Pavement Types	PCI_2014	PCI_2015
Portland Cement	• PCI_2013 • IRI_Index • Age_2013 • Annual_18_KIPS • Speed_Limit • Pavement_Width • Average_Daily_Traffic	• PCI_2013 • IRI_Index • Average_Daily_Traffic • Age_2013
Composite	• PCI_2013 • IRI_Index • Pavement_Width • Reconstruct_18_KIPS • Accu_KIPS_Since_Resurfacing • Number_of_Lanes • Median • Age_2013	• PCI_2013 • Median • IRI_Index • DaysTempChange_2013 • Speed_Limit
Asphalt Cement	• PCI_2013 • RS_In2014 • Annual_18_KIPS • Average_Daily_Traffic • Average_Daily_Trucks	• PCI_2013 • RS_In2015 • Annual_18_KIPS • Average_Daily_Traffic • Friction_Value • Average_Daily_Trucks

Note that for the portland cement pavement type, PCI_2013, IRI_Index, Average_Daily_Traffic, and Age_2013 are the common key variables in predicting PCI for both 2014 and 2015. For the composite pavement type, PCI_2013, IRI_Index, Median (absent/present), and Speed_Limit are the same for the two predictive models. Further, for the asphalt cement pavement type, PCI_2013, Annual_18_KIPS, Average_Daily_Traffic, and Average_Daily_Trucks are the important variables across the two models.

The analysis using Watson Analytics reveals that a ML approach is a viable approach to predicting PCI because it identifies the key input variables (as shown in Table 5) for three different pavement types. The analysis also shows that it is possible to predict 2014 and 2015 PCI values using 2013 PCI readings and thus eliminate the need to measure PCI every year.

It is recommended that this analysis be repeated in the future with different data sets to ensure its generalizability and validity.