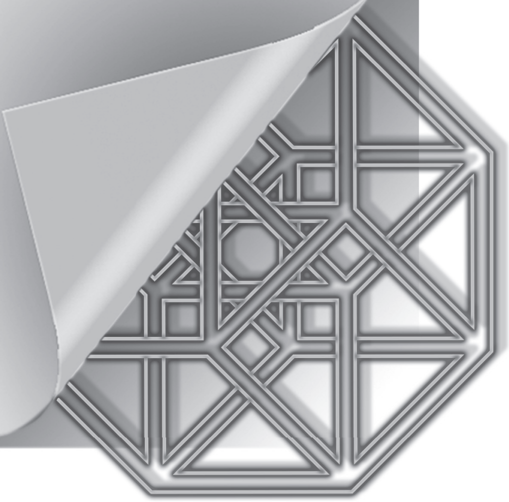
Fundamentals of Predictive Analytics with JMP®

Second Edition

Ron Klimberg · B. D. McCullough

Student Exercises



sas.com/books

The correct bibliographic citation for this manual is as follows: Klimberg, Ron and B.D. McCullough. 2016. *Fundamentals of Predictive Analytics with JMP®, Second Edition*. Cary, NC: SAS Institute Inc.

Fundamentals of Predictive Analytics with JMP®, Second Edition

Copyright © 2016, SAS Institute Inc., Cary, NC, USA

ISBN 978-1-62959-856-7 (Hard copy)
ISBN 978-1-62960-801-3 (EPUB)
ISBN 978-1-62960-802-0 (MOBI)
ISBN 978-1-62960-803-7 (PDF)

All Rights Reserved. Produced in the United States of America.

For a hard copy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government License Rights; Restricted Rights: The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication, or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a), and DFAR 227.7202-4, and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

December 2016

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

SAS software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For license information about third-party software distributed with SAS software, refer to **http://support.sas.com/thirdpartylicenses**.

Chapter 2: Exercises

1. Using the hmeq.jmp file and the variable Loan. Are there any outliers on the high or lower end?
2. Using the hmeq.jmp file and the variable Mortgage. Are there any outliers on the high or lower end?
3. Using the hmeq.jmp file, show the relationship between Mortgage and Loan. Discuss it.
4. Using the hmeq.jmp file, show the relationship between Mortgage and Reason. Discuss it.
5. Using the hmeq.jmp file, show the relationship between Default and Reason. Discuss it.
6. Using the Promotion_new.jmp file, and the variable Combine. Are there any outliers on the high or lower end?
7. Using the Promotion_new.jmp file, show the relationship between Race and Position. Discuss it.
8. Using the Promotion_new.jmp file, show the relationship between Race and Combine. Discuss it.
9. Using the Promotion_new.jmp file, show the relationship between Oral, Written and Combine. Discuss it.

Chapter 4: Exercises

1. Perform descriptive statistics and create some relevant graphs using the salesperfdata.jmp data set.
2. Perform descriptive statistics and create some relevant graphs using the churn.jmp data set.
3. Perform descriptive statistics and create some relevant graphs using the StateGDP2008.jmp data set.
4. Perform descriptive statistics and create some relevant graphs using the PublicUtilities.jmp data set.
5. Perform descriptive statistics and create some relevant graphs using the Kuiper.jmp data set.

Chapter 5: Exercises

1. Using the Countif.xls file, develop a regression model to predict Salary by using all the remaining variables. Use $\alpha = 0.05$. Evaluate this model—perform all the tests. Run a stepwise model and evaluate it.
2. Using the hmeq.jmp file, develop the best model you can to predict loan amount. Evaluate each model and use $\alpha = 0.05$.
3. Using the Promotion_new.jmp file, develop each model, evaluate it, and use $\alpha = 0.05$:
   a. Develop a model to predict Combine score, using the variable Position.
   b. Create an indicator variable for the variable Position, and use the new indicator variable Captain in a regression model to predict Combine.
   c. Create a new variable, called Position_2, which assigns 1 to Captain or −1 otherwise (leave blank as *missing*). Now run a regression model, using this new variable, Position_2, to predict Combine.
   d. What is the differences and similarities among parts a, b, and c?
   e. Develop a model using Race and Position to predict Combine.
4. Using the Countif.xls file, address the following:
   a. Is there significant interaction between Major and Gender and the average salary?
   b. Create a new variable, GPA_greater_3, which is equal to High if GPA $\geq$ 3; otherwise, it is equal to Low. See how this new variable, GPA_greater_3 has different salary means.
   c. Continuing with part b, test to see whether there is significant interaction between GPA_greater_3 and Major and in average Salary.
5. Using the Promotion_new.jmp file, test to see whether there is significant interaction with Race and Position and the average Combine score.
6. Using the Titanic_Passengers_new.jmp file, test to see whether there is significant interaction between whether a passenger survived and passenger class and the passengers' average age.

Chapter 6: Exercises

1. Consider the logistic regression for the toy data set, where π is the probability of passing the class:

$$\log\left[\frac{\hat{\pi}}{1-\hat{\pi}}\right] = 25.60188 - 0.363761\,\mathrm{MidtermScore}$$

   Consider two students, one who scores 67% on the midterm and one who scores 73% on the midterm. What are the odds that each fails the class? What is the probability that each fails the class?

   Consider the first logistic regression for the Churn data set, the one with 10 independent variables. Consider two customers, one with an international plan and one without. What are the odds that each churns? What is the probability that each churns?

2. You have already found that the interaction term **IntlPlanMins** significantly improves the model. Find another interaction term that does so.
3. Without deriving new variables such as **CustServ** or creating interaction terms such as **IntlPlanMins**, use a stepwise method to select variables for the Churn data set. Compare your results to the bivariate method used in the chapter; pay particular attention to the fit of the model and the confusion matrix.
4. Use the Freshmen1.jmp data set and build a logistic regression model to predict whether a student returns. Perhaps the continuous variables **Miles from Home** and **Part Time Work Hours** do not seem to have an effect. See whether turning them into discrete variables makes a difference. (*In essence*, turn **Miles from Home** into some dummy variables, such as 0–20 miles, 21–100 miles, or more than 100 miles.)

Chapter 7: Exercises

1. Use the PublicUtilities.jmpdata set. Run a regression to predict **return** using all the other variables. Run a PCA and use only a few principal components to predict **return** (remember not to include **return** in the variables on which the PCA is conducted).
2. Use the MassHousing.jmp data set. Run a regression to predict market value (**mvalue**) using all the other variables. Run a PCA and use only a few principal components to predict **mvalue.** (Remember not to include **mvalue** in the variables on which the PCA is conducted.)

Chapter 8: Exercises

1. For the Mass Housing data set, **MassHousing.jmp**, explore the effects of using Adaptive Estimation, as well as various form of Validation. Already you have used both the LASSO and the elastic net without Adaptive Estimation and with **AICc** validation. Now try it with **Adaptive Estimation** and **AICc** validation. Then try it without **Adaptive Estimation** and with **KFold** validation, and so on.
2. Do exercise 1 above, except use the Churn data set from Chapter 6 on Logistic Regression. Be sure to change the **Distribution** from **Normal** to **Binomial**.
3. Using the Sales Performance data set, **Salesperfdata.xls**, run a regression and stepwise regression. Then use LASSO and elastic net approaches. Compare the models. Which is best?
4. Using the Financial data set, Financial.jmp, run a regression and stepwise regression. Then use LASSO and elastic net approaches. Compare the models. Which is best?
5. Using the Freshman data set, Freshman.jmp, run a regression and stepwise regression. Then use LASSO and elastic net approaches. Compare the models. Which is best?

Chapter 9: Exercises

1. Use hierarchical clustering on the Public Utilities data set. Be sure to use the company name as a label. Use all six methods (for example, Average, Centroid, Ward, Single, Complete, and Fast Ward), and run each with the data standardized. How many clusters does each algorithm produce?
2. Repeat exercise 1, this time with the data not standardized. How does this affect the results?
3. Use hierarchical clustering on the Freshmen1.jmp data set. How many clusters are there? Use this number to perform a *k*-means clustering, (Be sure to try several choices of *k* near the one that is indicated by hierarchical clustering.) Note that *k*-means will not permit ordinal data. Based on the means of the clusters for the final choice of *k*, try to name each of the clusters.
4. Use *k*-means clustering on the churn data set. Try to name the clusters.

Chapter 10: Exercises

1. Build a classification tree on the churn data set. Remember that you are trying to predict churn, so focus on nodes that have many churners. What useful insights can you make about customers who churn?
2. After building a tree on the churn data set, use the **Column Contributions** to determine which variables might be important. Could these variables be used to improve the Logistic Regression developed in Chapter Five?
3. Build a Regression Tree on the Masshousing.jmp data set to predict market value.

Chapter 11: Exercises

1. Is there any point in extending *k* beyond 10 in the case of the glass data? Set the maximum value of *k* at 20 and then at 30.
2. Are the results for the glass data affected by the scaling? Standardize all the variables and apply K Nearest Neighbors.
3. Run a linear regression problem of your choice and compare it to K Nearest Neighbors.

Chapter 12: Exercises

1. Investigate whether the **Robust** option makes a difference. For the Kuiper data that has been used in this chapter (don't forget to drop some observations that are outliers!), run a basic model 20 times (for example, the type in Table 12.4). Run it 20 more times with the **Robust** option invoked. Characterize any differences between the results. For example, is there less variability in the R^2? Is the difference between training R^2 and validation R^2 smaller? Now include the outliers, and redo the analysis. Has the effect of the **Robust** option changed?
2. For all the analyses of the Kuiper data in this chapter, ten observations were excluded because they were outliers. Include these observations and rerun the analysis that produced one of the tables in this chapter (for example, Table 12.4). What is the effect of including these outliers? You will have to set the seed for the random number generator, and use the same seed twice: once when you run the model with the excluded observations, and again after you include the observations. If you don't do this, then you can't be sure whether the differences are due to the inclusion/exclusion of observations or the different random numbers!
3. For the neural net prediction of the binary variable MedPrice, try to find a suitable model by varying the architecture and changing the options.
4. Develop a neural network model for the Churn data.
5. In Chapter 10, you developed trees in two cases: a classification tree to predict whether students return and a regression tree to predict GPA. Develop a neural network model for each case.
6. As indicated in the text, sometimes rescaling variables can improve the performance of a neural network model. Rescale the variables for an analysis presented in the chapter (or in the exercises), and see whether the results improve.

Chapter 13: Exercises

1. Without using a Validation column, run a logistic regression on the Titanic data and compare to the results in this chapter.
2. Can you improve on the results in Figure 1?
3. How high can you get the RSquare in Mass Housing example used in Figure 4?
4. Without using a validation column, apply logistic regression, bootstrap forests, and boosted tree to the Churn data set.
5. Use a validation sample on boosted regression trees with Mass Housing. How high can you get the RSquared on the validation sample? Compare this to your answer for question (3).
6. Use a validation sample, and apply logistic regression, bootstrap forests, and boosted trees to the Churn data set. Compare this answer to your answer for question (4).

Chapter 14: Exercises

1. Create 30 columns of random numbers and use stepwise regression to fit them (along with S&P500) to the McDonalds return data.

   To create the 30 columns of random normal, first copy the McDonalds72 data set to a new file (say, McDonalds72-A). Open the new file and delete the 30 columns of “stock” data. Select **Cols ▶ New Columns**. Leave the **Column prefix** as **Column** and for **How many columns to add?** Enter 30. Under **Initial Data Values**, select **Random**, and then select **Random Normal**, and click **OK**.

   After running the stepwise procedure, take note of the RSquared and the number of “significant” variables added to the regression. Repeat this process 10 times. What are the highest and lowest R^2 that you observe? What are the highest and lowest number of statistically significant random variables added to the regression?

   a. Use the churn data set and run a logistic regression with three independent variables of your choosing. Create Lift and ROC charts, as well as a confusion matrix. Now do the same again, this time with six independent variables of your choosing. Compare the two sets of charts and confusion matrices.

   b. Use the six independent variables from the previous exercise and develop a neural network for the churn data. Compare this model to the logistic regression that was developed in that exercise.

2. Use the Freshmen1.jmp data set. Use logistic regression and classification trees to model the decision for a freshman to return for the sophomore year. Compare the two models using Lift and ROC charts, as well as confusion matrices.

Chapter 15: Exercises

1. In the aircraft_incidents.jmp file is data for airline incidents were retrieved on November 20th, 2015 from http://www.ntsb.gov/_layouts/ntsb.aviation/Index.aspx. For the Final Narrative variable, use the Text Explorer to produce a DTM by phrasing and terming. Create a Word Cloud.
2. As in problem 1, similarly produce a DTM by phrasing, terming, and create a Word Cloud except for the variable Narrative Cause.
3. In the file Nicardipine.jmp is data from adverse events from this drug. For the Reported Term for the Adverse Event variable, use the Text Explorer to produce a DTM by phrasing and terming. Create a Word Cloud.
4. In the Airplane_Crash_Reports.jmp file is one variable, NTSB Narrative that describes that summarizes the crash report. For this variable, use the Text Explorer to produce a DTM by phrasing and terming. Create a Word Cloud.
5. In the F DA_Enforcement_Actions.jmp file, the variable Citation Description describes the violation. For this variable, use the Text Explorer to produce a DTM by phrasing and terming. Create a Word Cloud.
6. The traffic-violation_jun2015.jmp is similar to the file used in the chapter except that the data is for June 2015 only. For the variable Description, use the Text Explorer to produce a DTM by phrasing and terming. Create a Word Cloud. How does this compare to data for December 2014?
7. Perform Latent Semantic Analytics, Topic Analysis and Cluster Analysis on the DTM you produced in Problem 1.
8. Perform Latent Semantic Analytics, Topic Analysis and Cluster Analysis on the DTM you produced in Problem 2.
9. Perform Latent Semantic Analytics, Topic Analysis and Cluster Analysis on the DTM you produced in Problem 3.
10. Perform Latent Semantic Analytics, Topic Analysis and Cluster Analysis on the DTM you produced in Problem 4
11. Perform Latent Semantic Analytics, Topic Analysis and Cluster Analysis on the DTM you produced in Problem 5.
12. Perform Latent Semantic Analytics, Topic Analysis and Cluster Analysis on the DTM you produced in Problem 6. How does this compare to data for December 2014?
13. Similar to what we did in the Chapter, create a predictive model for violation type. How does this compare to data for December 2014?

Chapter 16: Exercises

1. Identify other opportunities for increasing Coke sales, or for using Coke to spur sales of other goods.
2. For the data in Table 16.1, compute Confidence and Lift for X = {bread} and Y= {jelly, peanut butter} by hand. Then check your answer by using JMP.
3. Calculate the Support of X = {bread} and Y = {jelly, peanut butter}.
4. Analyze the GroceryPurchases.jmp data and find some actionable rules.