Adv. Methods

Module 7 – Advanced Analytics - Theory and Methods Part II

EMC^2 PROVEN PROFESSIONAL

Module 7: Advanced Analytics – Theory and Methods – Part II

Upon completion of this module, you should be able to:

- Examine analytic needs and select an appropriate technique based on business objectives; initial hypotheses; and the data's structure and volume
- Apply some of the more commonly used methods in Analytics solutions
- Explain the algorithms and the technical foundations for the commonly used methods
- Explain the environment (use case) in which each technique can provide the most value
- Use appropriate diagnostic methods to validate the models created
- Use R and in-database analytical functions to fit, score and evaluate models

What Kind of Problem do I Need to Solve?
How do I Solve it? ***<this module focuses on Regression>***

The Problem to Solve	The Category of Techniques	Covered in this Course
I want to group items by similarity. I want to find structure (commonalities) in the data	Clustering	K-means clustering
I want to discover relationships between actions or items	Association Rules	Apriori
I want to determine the relationship between the outcome and the input variables	Regression	Linear Regression Logistic Regression
I want to assign (known) labels to objects	Classification	Naïve Bayes Decision Trees
I want to find the structure in a temporal process I want to forecast the behavior of a temporal process	Time Series Analysis	ACF, PACF, ARIMA
I want to analyze my text data	Text Analysis	Regular expressions, Document representation (Bag of Words), TF-IDF

Module 7: Advanced Analytics – Theory and Methods

Lesson: Linear Regression

During this lesson the following topics are covered:

- General description of regression models
- Technical description of a linear regression model
- Common use cases for the linear regression model
- Interpretation and scoring with the linear regression model
- Diagnostics for validating the linear regression model
- The Reasons to Choose (+) and Cautions (-) of the linear regression model

Regression

- Regression focuses on the relationship between an outcome and its input variables.
  - In other words, we don't just predict the outcome, we also have a sense of how changes in individual drivers affect the outcome.
- The outcome can be continuous or discrete.
  - When it's discrete, we are predicting the probability that the outcome will occur.

Example Questions:

- I want to predict the life time value (LTV) of this customer (and understand what drives LTV).
- I want to predict the probability that this loan will default (and understand what drives default).

- **Our examples: Linear Regression, Logistic Regression**

Linear Regression -What is it?

- Used to estimate a continuous value as a linear (additive) function of other variables
  - Income as a function of years of education, age, gender
  - House price as function of median home price in neighborhood, square footage, number of bedrooms/bathrooms
  - Neighborhood house sales in the past year based on unemployment, stock price etc.

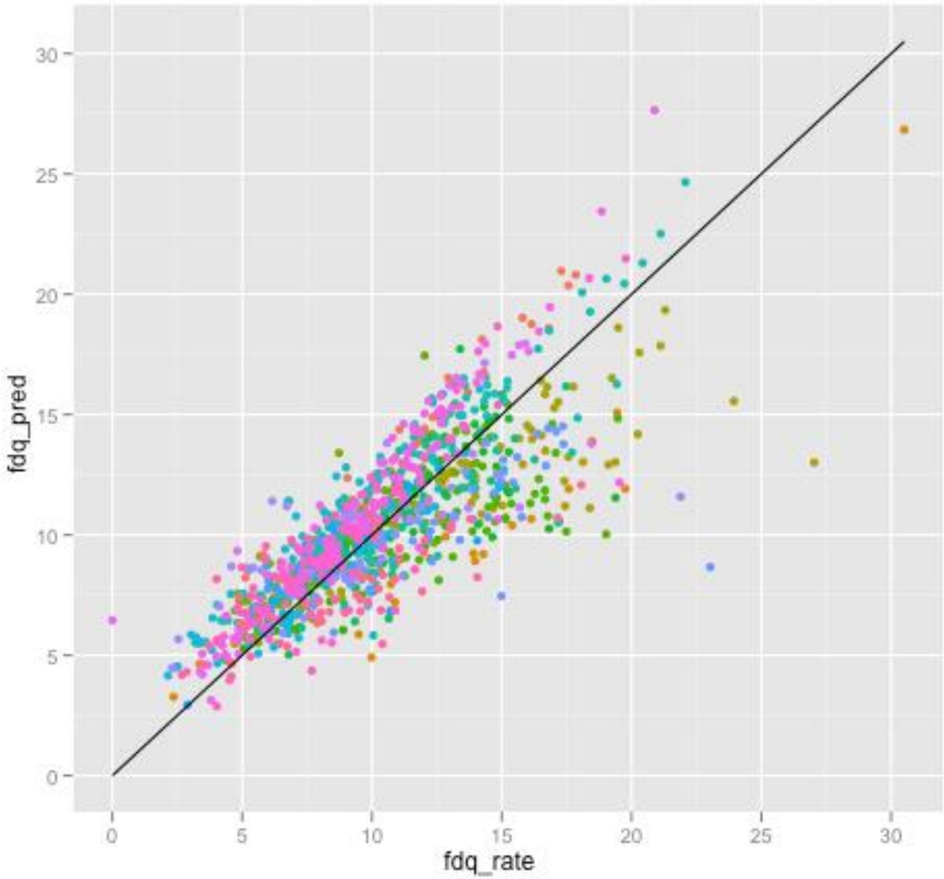
- Input variables can be continuous or discrete.
- Output:
  - A set of coefficients that indicate the relative impact of each driver.
  - A linear expression for predicting outcome as a function of drivers.

Linear Regression - Use Cases

- The preferred method for almost any problem where we are predicting a continuous outcome
  - Try this first; if it fails, then try something more **complicated**
- Examples:
  - Customer lifetime value
  - Home value
  - Loss given default on loan
  - Income as a function of demographics

Example: Predict Mortgage Foreclosure/Delinquency Rates

fdq_rate = -0.9 + 0.66 CurrentUnemp + 1.06 ChgInUnem1yr + 0.22 hicost_mort_rate



Technical Description

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots.$$

- Solve for the b_i
  - Ordinary Least Squares
    - storage quadratic in number of variables
    - must invert a matrix
- Categorical variables are expanded to a set of indicator variables, one for each possible value.

Representing Categorical Variable

$$\text{income} = b_0 + b_1\text{age} + b_2\text{yearsOfEducation} + b_3\text{gender} + b_4\text{state}$$

- *State* is a categorical variable: 50 possible values.
- Expand it to 49 indicator (0/1) variables:
  - The remaining level is the "default level“
  - This is done automatically by standard packages
- *Gender* is categorical, too, but binary
  - so one variable: *genderMale*, which is 0 for females

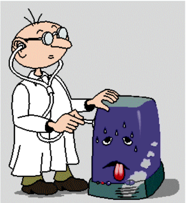
What do the Coefficients b_i Mean?

- Change in y as a function of unit change in x_i
  - all other things being equal
- Example: income in units of $10K, years in age, $b_{age} = 2$
  - For the same gender, years of education, and state of residence, a person's income increases by 2 units (20K)for every year older

- Standard packages also report the significance of the b_i: probability that, in reality, $b_i = 0$
  - b_i "significant" if P($b_i = 0$) is small

EMC[2] PROVEN PROFESSIONAL

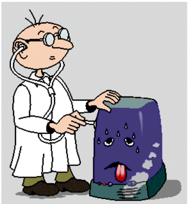
Diagnostics



- Hold-out data
  - Does the model predict well on data it hasn't seen?
- N-fold cross-validation
  - Partition the data into N groups.
  - Fit N models, holding out each group, and calculate the residuals on the group.
  - Estimated prediction error is the average over all the residuals.
- R^2 : The fraction of the variance in the output variable that the model can explain.
  - It is also the square of the correlation between the true output and the predicted output. You want it close to 1.

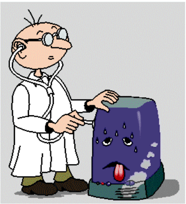
EMC^2 PROVEN PROFESSIONAL

Diagnostics (Continued)



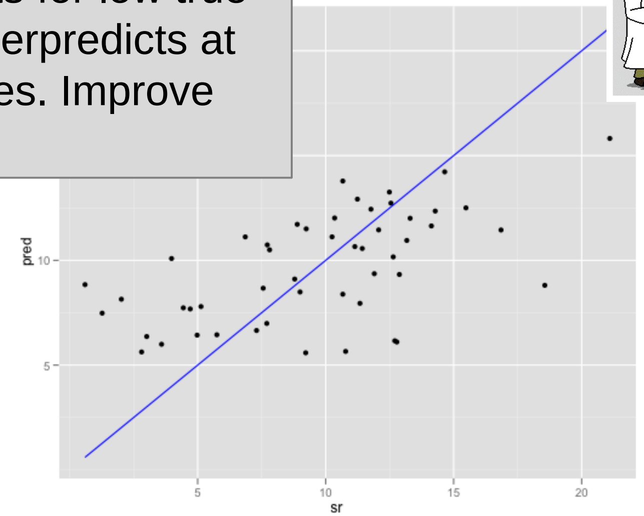
- Sanity check the coefficients
  - Do the signs make sense? Are the coefficients excessively large?
    - Wrong sign is an indication of correlated inputs, but doesn't necessarily affect predictive power.
    - Excessively large coefficient magnitudes may indicate strongly correlated inputs; you may want to consider eliminating some variables, or using regularized regression techniques.
      - Ridge, Lasso
    - Infinite magnitude coefficients could indicate a variable that strongly predicts a subset of the output (and doesn't predict well on the rest).
      - Plot output vs. this input, and see if you should segment the data before regressing.

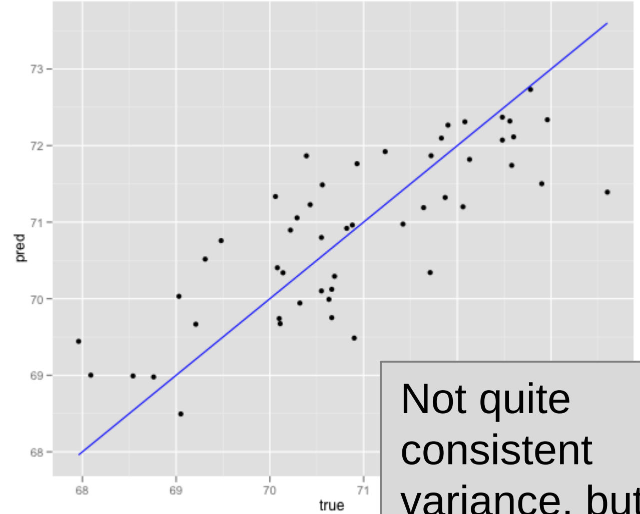
Diagnostics (Continued)



- **Plot it!**
  - Prediction vs. true outcome
- Look for:
  - Systematic over/under prediction
  - Non-consistent variance
    - The data cloud should be symmetric about the line of true prediction
  - Glaring outliers
- You will see other diagnostic plots in the lab

Overpredicts for low true values, underpredicts at higher values. Improve the model.





Not quite consistent variance, but much better.

Linear Regression - Reasons to Choose (+) and Cautions (-)

Reasons to Choose (+)	Cautions (-)
Concise representation (the coefficients)	Does not handle missing values well
Robust to redundant variables, correlated variables Lose some explanatory value	Assumes that each variable affects the outcome linearly and additively Variable transformations and modeling variable interactions can alleviate this A good idea to take the log of monetary amounts or any variable with a wide dynamic range
Explanatory value Relative impact of each variable on the outcome	Can't handle variables that affect the outcome in a discontinuous way Step functions
Easy to score data	Doesn't work well with discrete drivers that have a lot of distinct values For example, ZIP code

Check Your Knowledge

Your Thoughts?

1. How is the measure of significance used in determining the explanatory value of a driver with linear regression models?
2. Detail the challenges with categorical values in linear regression model.
3. Describe N-Fold cross validation method used for diagnosing a fitted model.
4. List two use cases of linear regression models.
5. List and discuss two standard sanity checks that you will perform on the coefficients derived from a linear regression model.

Module 7: Advanced Analytics – Theory and Methods – Part II

Lesson: Linear Regression - Summary

During this lesson the following topics were covered:

- General description of regression models
- Technical description of a linear regression model
- Common use cases for the linear regression model
- Interpretation and scoring with the linear regression model
- Diagnostics for validating the linear regression model
- The Reasons to Choose (+) and Cautions (-) of the linear regression model

Lab Exercise 6: Linear Regression

This Lab is designed to investigate and practice Linear Regression.

After completing the tasks in this lab you should be able to:

- Use R functions for Linear Regression (Ordinary Least Squares – OLS)
- Predict the dependent variables based on the model
- Investigate different statistical parameter tests that measure the effectiveness of the model

Lab Exercise 6: Linear Regression - Workflow

1. Set Working directory
2. Use random number generators to create data for the OLS Model
3. Generate the OLS model using R function "lm"
4. Print and visualize the results and review the plots generated
5. Generate Summary Outputs
6. Introduce a slight non-linearity and test the model
7. Perform In-database Analysis of Linear Regression

Module :7 Advanced Analytics – Theory and Methods – Part II

Lesson: Logistic Regression

During this lesson the following topics are covered:

- Technical description of a logistic regression model
- Common use cases for the logistic regression model
- Interpretation and scoring with the logistic regression model
- Diagnostics for validating the logistic regression model
- Reasons to Choose (+) and Cautions (-) of the logistic regression model

EMC^2 PROVEN PROFESSIONAL

Logistic Regression

- Used to estimate the probability that an event will occur as a function of other variables
  - The probability that a borrower will default as a function of his credit score, income, the size of the loan, and his existing debts
- Can be considered a classifier, as well
  - Assign the class label with the highest probability

- Input variables can be continuous or discrete
- Output:
  - A set of coefficients that indicate the relative impact of each driver
  - A linear expression for predicting the log-odds ratio of outcome as a function of drivers. (Binary classification case)
    - Log-odds ratio easily converted to the probability of the outcome

Logistic Regression Use Cases

- The preferred method for many binary classification problems:
  - Especially if you are interested in the probability of an event, not just predicting the "yes or no“
  - Try this first; if it fails, then try something more complicated
- Binary Classification examples:
  - The probability that a borrower will default
  - The probability that a customer will churn
- Multi-class example
  - The probability that a politician will vote yes/vote no/not show up to vote on a given bill

Logistic Regression Model - Example

$$\text{default} = f(\text{creditScore},\ \text{income},\ \text{loanAmt},\ \text{existingDebt})$$

- Training data: default is 0/1
  - default=1 if loan defaulted
- The model will return the probability that a loan with given characteristics will default
- If you only want a "yes/no" answer, you need a threshold
  - The standard threshold is 0.5

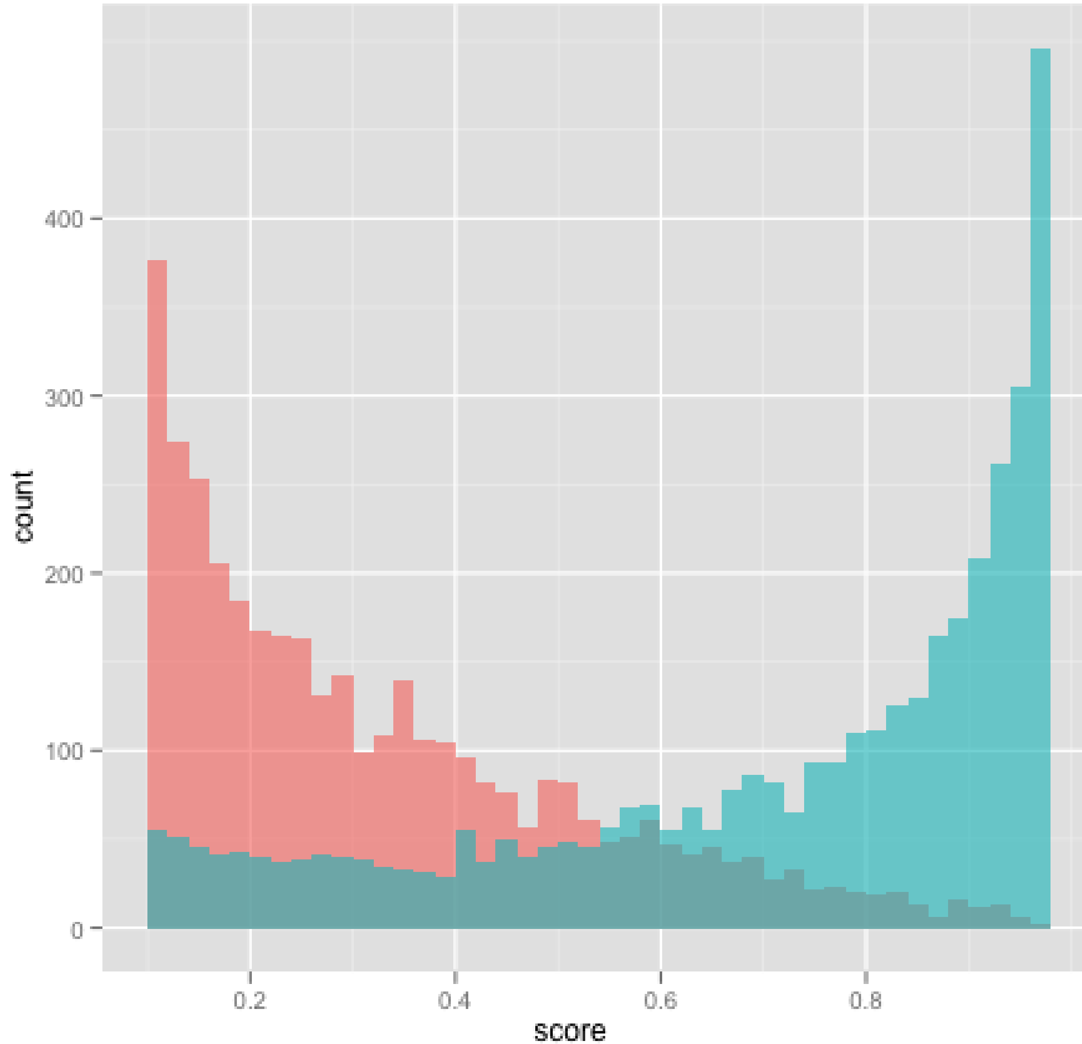
Logistic Regression- Visualizing the Model

Overall fraction of default: ~20%

Logistic regression returns a score that estimates the probability that a borrower will default

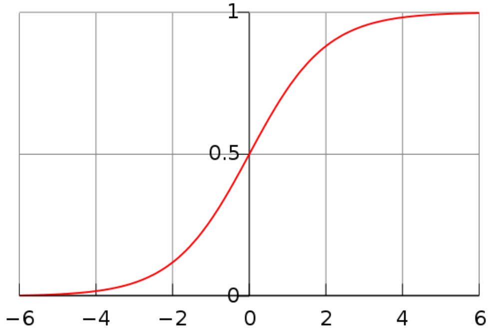
The graph compares the distribution of defaulters and non-defaulters as a function of the model's predicted probability, for borrowers scoring higher than 0.1

Blue=defaulters



Technical Description (Binary Case)

$$\ln \frac{P(y=1)}{1-P(y=1)} = b_0 + b_1 x_1 + b_2 x_2 \ldots$$



- y=1 is the case of interest: 'TRUE'
- LHS is called logit(P(y=1))
  - hence, "logistic regression"
- logit(P(y=1)) is inverted by the sigmoid function
  - standard packages can return probability for you
- Categorical variables are expanded as with linear regression
- Iterative, not closed form solution
  - "Iteratively re-weighted least squares"

What do the Coefficients b_i *Mean?*

- Invert the logit expression:

$$\frac{P(y=1)}{1-P(y=1)} = \exp(\sum_{j=0}^{K} b_j x_j)$$

$$= \prod_{j=0}^{K} \exp(b_j x_j)$$

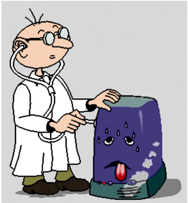
- $\exp(b_j)$ tells us how the odds-ratio of y=1 changes for every unit change in x_j
- Example: $b_{creditScore}$ = -0.69
  - $\exp(b_{creditScore})$ = 0.5 = 1/2
  - for the same income, loan, and existing debt, the odds-ratio of default is halved for every point increase in credit score
- Standard packages return the significance of the coefficients in the same way as in linear regression

An Interesting Fact About Logistic Regression

"The probability mass equals the counts"

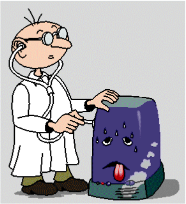
- If 13% of our loan risk training set defaults
  - The sum of all the training set scores will be 13% of the number of training examples
- If 40% of applicants with income < $50,000 default
  - The sum of all the training set scores of people in this income category will be 40% of the number of examples in this income category

Diagnostics



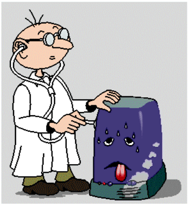
- Hold-out data:
  - Does the model predict well on data it hasn't seen?
- N-fold cross-validation: Formal estimate of generalization error
- "Pseudo-R^2" : 1 – (deviance/null deviance)
  - Deviance, null deviance both reported by most standard packages
  - The fraction of "variance" that is explained by the model
  - Used the way R^2 is used

Diagnostics (Cont.)



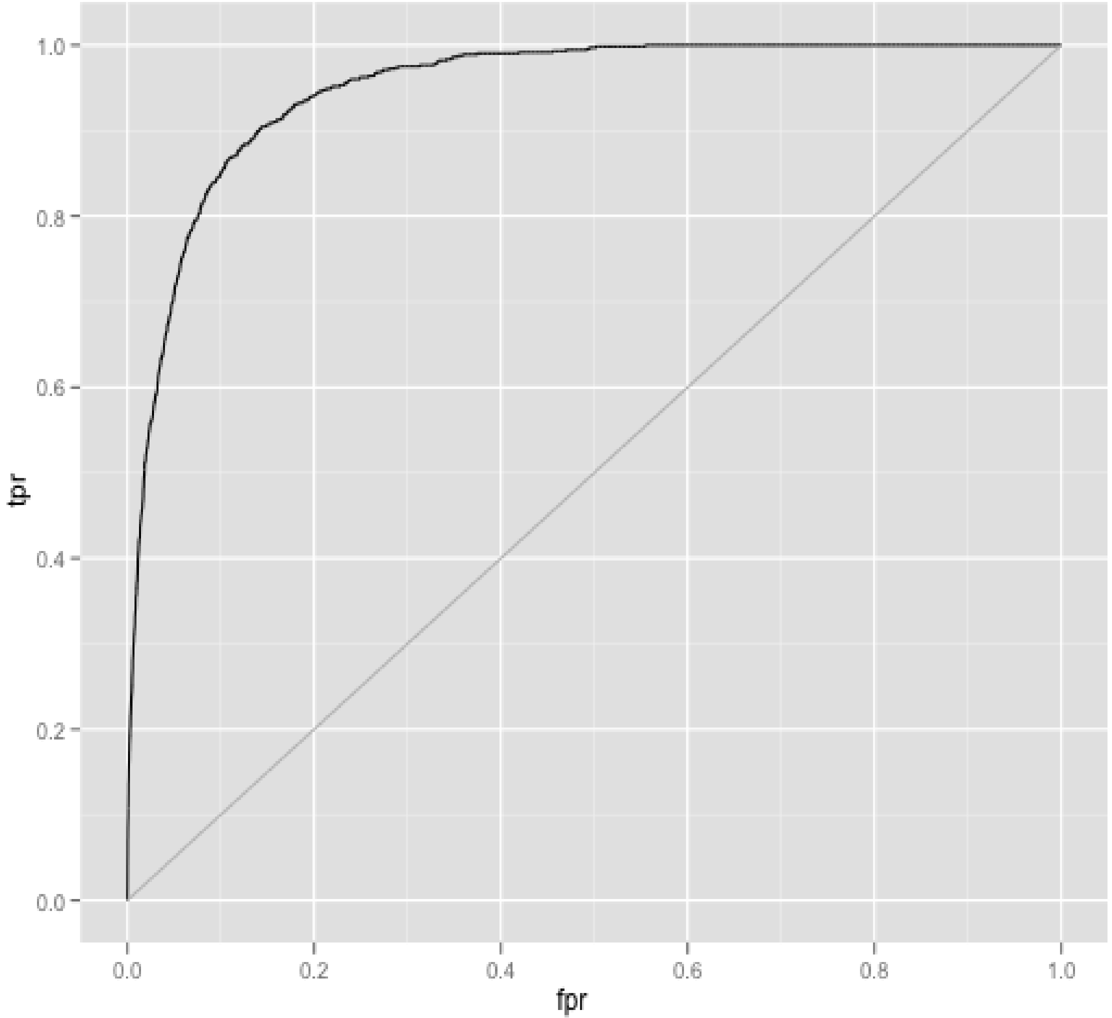
- Sanity check the coefficients
  - Do the signs make sense? Are the coefficients excessively large?
    - Wrong sign is an indication of correlated inputs, but doesn't necessarily affect predictive power.
    - Excessively large coefficient magnitudes may indicate strongly correlated inputs; you may want to consider eliminating some variables**, or using regularized regression techniques.**
      - Unfortunately, regularized logistic regression is not standard.
    - Infinite magnitude coefficients could indicate a variable that strongly predicts a subset of the output (and doesn't predict well on the rest).
      - Try a Decision Tree on that variable, to see if you should segment the data before regressing.

Diagnostics: ROC Curve



$$\text{FPR} = \frac{\#\ \text{false positives}}{\text{all negatives}}$$

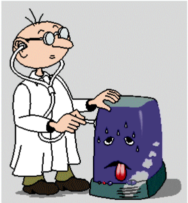
$$\text{TPR} = \frac{\#\ \text{true positives}}{\text{all positives}}$$

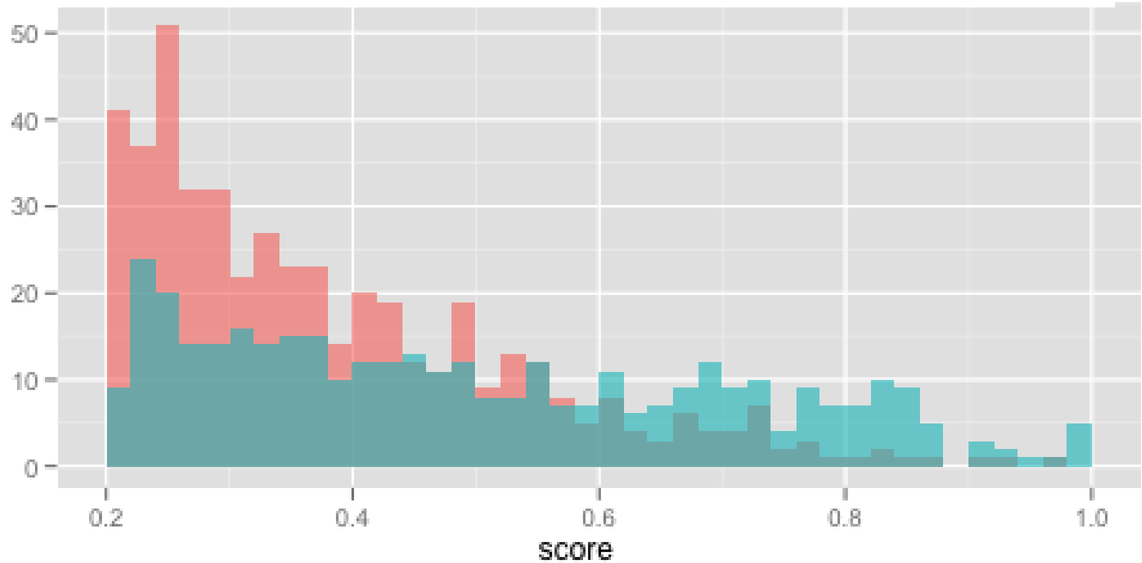


Area under the curve (AUC) tells you how well the model predicts. (Ideal AUC = 1)

For logistic regression, ROC curve can help set classifier threshold

Diagnostics: Plot the Histograms of Scores





EMC^2 PROVEN PROFESSIONAL

Logistic Regression - Reasons to Choose (+) and Cautions (-)

Reasons to Choose (+)	Cautions (-)
Explanatory value: Relative impact of each variable on the outcome in a more complicated way than linear regression	Does not handle missing values well
Robust with redundant variables, correlated variables Lose some explanatory value	Assumes that each variable affects the log-odds of the outcome linearly and additively Variable transformations and modeling variable interactions can alleviate this A good idea to take the log of monetary amounts or any variable with a wide dynamic range
Concise representation with the the coefficients	Cannot handle variables that affect the outcome in a discontinuous way. Step functions
Easy to score data	Doesn't work well with discrete drivers that have a lot of distinct values For example, ZIP code
Returns good probability estimates of an event	
Preserves the summary statistics of the training data "The probabilities equal the counts"	

Check Your Knowledge

Your Thoughts?

1. What is a logit and how do we compute class probabilities from the logit?
2. How is ROC curve used to diagnose the effectiveness of the logistic regression model?
3. What is Pseudo R^2 and what does it measure in a logistic regression model?
4. How do you describe a binary class problem?
5. Compare and contrast linear and logistic regression methods.

Module 7: Advanced Analytics – Theory and Methods Part II

Lesson: Logistic Regression - Summary

During this lesson the following topics were covered:

- Technical description of a logistic regression model
- Common use cases for the logistic regression model
- Interpretation and scoring with the logistic regression model
- Diagnostics for validating the logistic regression model
- Reasons to Choose (+) and Cautions (-) of the logistic regression model

Lab Exercise 7: Logistic Regression

This Lab is designed to investigate and practice Logistic Regression.

After completing the tasks in this lab you should be able to:

- Use R functions for Logistic Regression – (*also known as Logit*)
- Predict the dependent variables based on the model
- Investigate different statistical parameter tests that measure the effectiveness of the model

Lab Exercise 7: Logistic Regression - Workflow

1. Set the Working Directory
2. Define the problem and review input data
3. Read in and Examine the Data
4. Build and Review logistic regression Model
5. Review and interpret the coefficients
6. Visualize the Model Using the Plot Function
7. Use relevel Function to re-level the Price factor with value 30 as the base reference
8. Plot the ROC Curve
9. Predict Outcome given Age and Income
10. Predict outcome for a sequence of Age values at price 30 and income at its mean
11. Predict outcome for a sequence of income at price 30 and Age at its mean
12. Use Logistic regression as a classifier