Machine Learning in Business
John C. Hull

Chapter 9
Model Interpretability

Why is model interpretability important?

- Users must understand a model to have confidence in it, know when it is appropriate, be aware of its biases, etc
- It is also important to be able to explain the predictions made by the model, e.g.,
  - Why was someone refused for a loan?
  - Why is house A worth more than house B
- The General Data Protection Regulation in the European Union requires model interpretability

Amusing Stories

- Hans: the horse that could do math
- Image recognition software to distinguish dogs from polar bears

White-box vs black-box models

- White-box models
  - *k*-nearest neighbors
  - Decision trees
  - Linear regression
- Black-box models
  - SVM
  - Neural networks
  - Ensemble models (e.g. random forests)

Linear Regression

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_mX_m$$

- The weights in a linear regression are easy to understand
- If the value of feature j changes by u the value of the estimate changes by b_ju
- The bias, a, is more difficult. It is the estimate when all features are zero. But zero values for the features might be impossible.
- A better way of expressing the model is

$$Y = a^* + b_1(X_1 - \bar{X}_1) + b_2(X_2 - \bar{X}_2) + \cdots + b_m(X_m - \bar{X}_m)$$

- The bias is then the estimate when all features have their average values

Calculating feature contributions in linear regression

- We can compare a currently observed feature value with the average feature value to determine the contribution of that feature to the total value.
- The sum of the contributions equals the difference between the current prediction and the prediction when all features have their average values
- Results for Iowa house price (Lasso model; first 4 features)

Feature	*House value*	*Average value*	*Feature weight*	*Contrib-ution ($)*
Lot area (sq. ft.)	15,000	10,249	0.3795	+1,803
Overall quality (1 to 10)	6.0	6.1	16,695	−1,669
Year built	1990	1972	134.4	+2,432
Year remodeled	1990	1985	241.2	+1,225

Feature Dependence

- Even in the Lasso model there is some dependence between features
- Total basement sq. ft. and first floor sq. ft. are not independent and it may not make sense to consider the effect of changing one without changing the other
- This is a problem in all models
- We might be able to group features that should be considered together. Sometimes a PCA is used to create uncorrelated features.

Logistic Regression

$$\text{Prob (Positive Outcome)} = \frac{1}{1 + \exp[-(a + b_1X_1 + b_2X_2 + \cdots + b_mX_m)]}$$

$$\text{Prob (Negative Outcome)} = \frac{\exp[-(a + b_1X_1 + b_2X_2 + \cdots + b_mX_m)]}{1 + \exp[-(a + b_1X_1 + b_2X_2 + \cdots + b_mX_m)]}$$

We can calculate the sensitivity of these to the feature values but the result is only good for small changes

For large changes we can use the formulas multiple times

Odds

- Odds of a positive result is

$$\exp[-(a + b_1X_1 + b_2X_2 + \cdots + b_mX_m)] \text{ to } 1 \quad \text{against}$$

or

$$\exp(a + b_1X_1 + b_2X_2 + \cdots + b_mX_m) \text{ to } 1 \quad \text{on}$$

$$\text{Probability} = \frac{1}{1 + \text{odds against}} = \frac{\text{odds on}}{1 + \text{odds on}}$$

If we are prepared to work we log(odds) we have linearity and can proceed as for linear regression

Black-box models

- Models must be re-run to determine the impact of the change in a feature value on a prediction
- In general there is non-linearity so that when changes are made to the feature values the sum of the contributions of the features does not equal the change in the prediction

Partial Dependence Plot

- The partial dependence plot is the expected prediction as a function of the value of a particular feature.
- The values of all features except the one under consideration are chosen randomly

Shapley Values

- Shapley values are a particular way of calculating feature contributions so that the sum of the contributions equals the change that is being explained
- They are based on the work of Lloyd Shapley in game theory

Example: Features are changed from "average" to "current values"

Feature 1 Value	Feature 2 Value	Feature 3 Value	Prediction
Average	Average	Average	100
Average	Average	Current	120
Average	Current	Average	125
Average	Current	Current	130
Current	Average	Average	110
Current	Average	Current	128
Current	Current	Average	137
Current	Current	Current	140

Consider all the sequences in which changes can happen and average the contributions

Sequence	*Feature 1 Contribution*	*Feature 2 Contribution*	*Feature 3 Contribution*
123	10	27	3
132	10	12	18
213	12	25	3
231	10	25	5
312	8	12	20
321	10	10	20
Average	10	18.5	11.5

Total contribution = 40 which is the total change in the prediction

Properties of Shapley values when used as contributions

- If a feature never changes the prediction, its contribution is zero.
- If two features are symmetrical in that they affect the prediction in the same way, they have the same contribution.
- For an ensemble model where predictions are the average of predictions given by several underlying models, the Shapley value is the average of the Shapley values for the underlying models.
- Calculation time increases exponentially with the number of features

LIME

- LIME tries to understand a black-box model by fitting a simpler model to data that is close to the currently observed data
- Procedure is:
  - Perturb feature values to get a samples
  - Run black-box model to get predictions for samples
  - Train an easy to interpret model such as linear regression or decision trees to fit the data set that is created from samples and predictions