Feature Engineering & Model Interpretation

- Hunter

Quiz

- ✗ How to convert Categorical to Numerical?
- ✗ If you have time series data how you handle it for modeling
- ✗ If you have transactional data how you handle it?
- ✗ How you interpret your model?

Topics

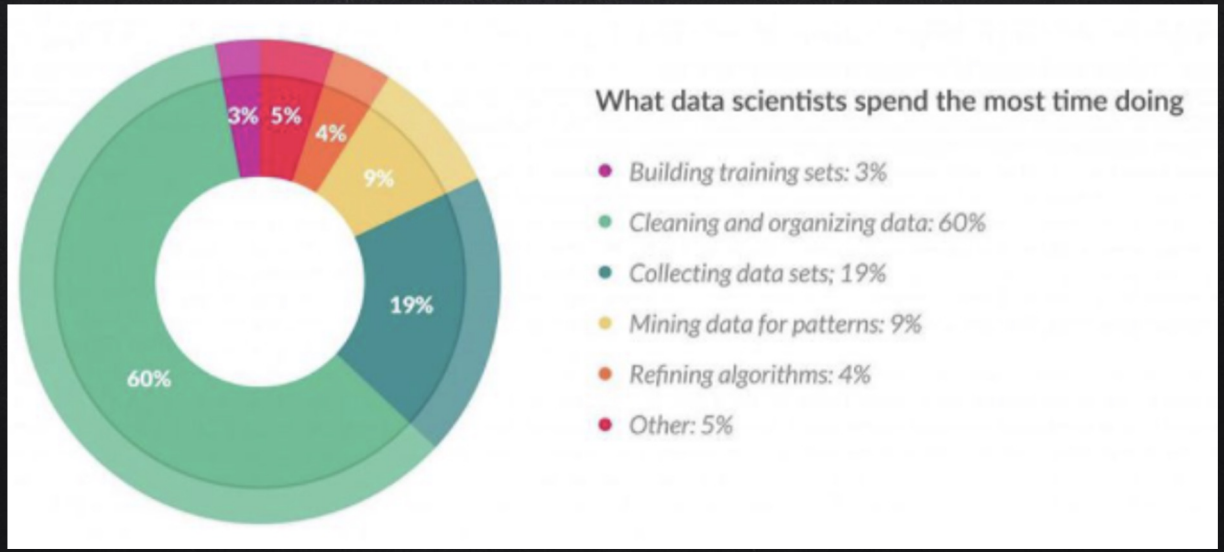
- ✗ Feature Engineering
  - Data Handling
  - Numerical transformation
  - Categorical to Numerical (grouping/encoding/transformation)
  - Features creation and data aggregation
- ✗ Model interpretation
  - Interpretable Models
  - Black Box Models

Feature Engineering

"Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data"

Data Decides ML Solution Quality

This is a survey in Forbes, data scientists spend 80% of their time on data preparation:



Source: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

Basic Data Handling

- ✗ Imputation – Handling missing, one of the most common problems you can encounter for machine learning

  Ignore, mean/median, fixed number, regression, tree based model, categorical imputation….

  Imputation Methods

- ✗ Outlier Handling

  Drop or Cap/Floor (1/99, 5/95)

  How to determine outliers?

  Q1–k*IQR, Q3+k*IQR. Using the interquartile multiplier value k=1.5

  mean – k*STD, mean + k*STD, k=3

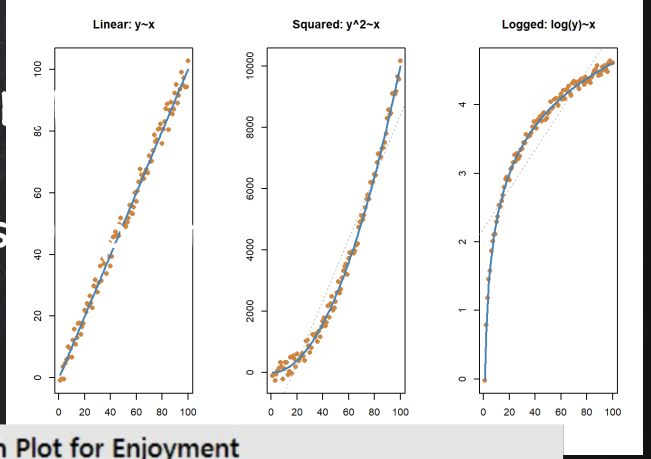
Numerical Transformati

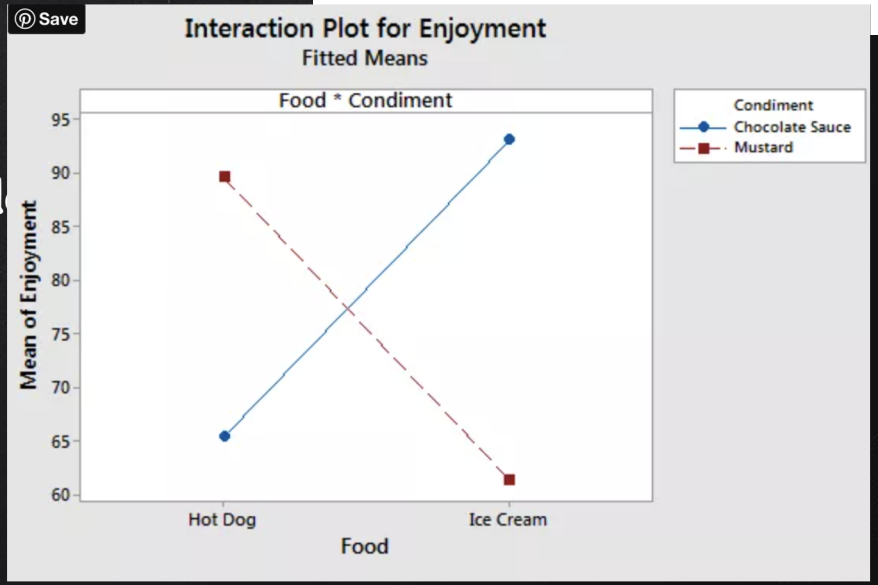
- ✗ Transform variables to meet parametric model as
  - Linearity

    Add additive terms, f(x) transformation
  - Random error

    f(y) transformation
- ✗ Transform variables for better mod
  - Spline terms transformation

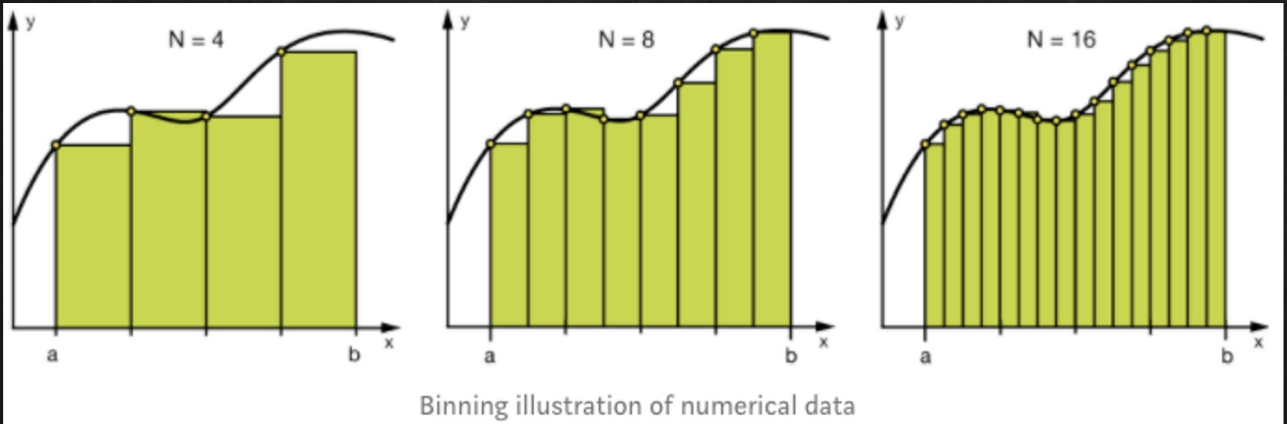
    Simple/Quadratic/Cubic
  - Interaction terms





Numerical Transformation – cont.

- Transform variables for model simplicity
  - WOE(weight of evidence) binning



Binning illustration of numerical data

Example after binning:

Value	Bin	WOE
0–500	Worst	–1.8
500–650	Fai	–0.567
650–750	Good	0.356
750 +	excellent	1.5

Categorical to Numerical

- ✗ One Hot Encoding
- ✗ Target Based Encoding
- ✗ Weight of Evidence (WOE)
- ✗ Leave-One-Out Encoding
- ✗ Ordinal Encoding
- ✗ Cardinality Reduction

MainPackage

One-Hot-Encoding

OneHot Encoding also called dummy coding in statistics. To build model all input should be numerical, the easiest way in statics to convert categorical variables to numerical is converting this categorical variable of n values into n-1 dummy variables (n-2 new columns)

Why n-1 ? Avoiding multicollinearity

One-hot-encoding VS Dummy encoding

Original	dummy_cat	dummy_pig	dummy_dog
Cat	1	0	0
Pig	0	1	0
Dog	0	0	1

One-Hot-Encoding tips and notes

- ★ Drop a level to avoid multicollinearity in a regression: If you are running a regression model, you need to drop the variable that shows multicollinearity with other variables. However, if you are running any tree-based algorithms, you should be fine.
- ★ Choose the level of credible count to be the base level: Again if I need to drop one which one to drop. The criterion is credibility. In a regression the dropped value of the categorical variable becomes the base level that other values will reference to. If the count of the dropped value is too few, it is not good to be the base level that other levels reference to. A better and common practice is to set the most frequent level, or a level that has sufficient number of observations.
- ★ Do not forget the missing values: data need has pretreatment with missing before encoding. Pay attention to a continuous variable has "NA", "0", "-99" or "-999".

Note: The disadvantage of the one-hot or dummy encoding is it could generates a very large sparse matrix if a categorical variable with many categories. Then it need grouping

Another disadvantage is not useful for ordinal data (ordinal VS nominal data)

Target Based Encoding

This method simply replaces each category with the mean target value for samples which have that category.

purpose	count	default	default_rate
credit_card	525	207	0.394286
major_purchase	115	49	0.426087
home_improvement	186	80	0.430108
debt_consolidation	1738	862	0.495972
other	318	172	0.540881
medical	30	19	0.633333
small_business	88	56	0.636364

Target Based Encoding Tips

- ★ This encoding method seriously plagued by overfitting, since it uses information about the target, this is target leakage.
- ★ A trick to overcome the issue, try to add tiny random noise to the new variable in order to prevent the model fit too perfectly for the training data. The next question will be how the tiny is? This is something should be tested
- ★ The random noise only need for build but not for test dataset

What is WOE

Weight of evidence (WOE) is a widely used technique in credit risk modeling. This transformation is aimed to get the maximum difference among the binned categories relating to the target variable.

- Formula of WOE and Information Value (IV) for binary target

  WOE = ln(% of non-events ÷ % of events)

  IV = ∑ (% of non-events – % of events) * WOE

- Formula of WOE for continuous target

  WOE = ln(% of target ÷ % records)

WOE Encoding

Similar like the target base encoder, the method try to map each categorical value to a numerical value, the value also anchored on target but use the WOE transformation to create the mapping value.

OCCUPATION_TYPE	COUNT	GOOD	BAD	GOOD%	BAD%	OCCUPATION_TYPE_WOE
Accountants	9813	474	9339	4.49	5.97	-0.28
High skill tech staff	11380	701	10679	6.64	6.82	-0.03
Managers	21371	1328	20043	12.58	12.80	-0.02
Core staff	27570	1738	25832	16.47	16.50	0.00
HR staff	563	36	527	0.34	0.34	0.01
NoData	96391	6278	90113	59.48	57.57	0.03
sum	**167088**	**10555**	**156533**	**100**	**100**	

Rules Related to WOE in Logistic Reg

- Each category (bin) should have at least 5% of the observations for a small dataset, for a big dataset at least 1% of the observations.
- Each category (bin) should be non-zero for both non-events and events, sometime maybe a good predictor
- The WOE should be distinct for each category. Similar groups should be aggregated. Should be set up certain rule with parameter tunable.
- The WOE should be monotonic, i.e. either growing or decreasing with the groupings. (Only applied to continuous variables)
- Missing values are binned separately.

WOE Encoding

- "Especially suitable for logistic regression: logistic regression fits a linear regression equation of predictors to predict the logit-transformed binary Goods/Bads target variable. The Logit transformation is simply the log of the odds. So by using the WOE-tranformed predictors in logistic regression, the predictors are coded to the same WOE scale, and the parameters in the linear logistic regression equation can be directly compared."
- Very good practice for categorical variable with high cardinality
- "Monotonic relationship with the target: between the target variable and the WOE-transformed variable. WoE transformation is strictly linear with respect to log odds of the response with the unity correlation."
- "There is no need to cap or floor the outliers and missing treatment, but pay attention to missing"

Leave One Out Encoding

Leave-one-out encoding is designed to overcome the target leakage.

When the producing the mapping of the numerical mean target value the row is left out of the calculation , therefore prevent the direct target leakage

Ordinal Encoding

When the categorical variable value have order instead pure nominal variables. This method assign level orders as numerical to the category values, or possible number orders by business intuition.

Level	EDU_1	EDU_2
Elementary	1	6
High School	2	12
2-year college	3	14
College graduate	4	18
Graduate degree	5	20

Cardinality Reduction

When a categorical variables have a lot of unique values, this variables has high cardinality, one step encoding will not fulfill the goal to understand the variable in your model or access the business intuition or impact through this variable, especially when this is a important variable. There is the need to reduce variable cardinality

Most time the reduction of cardinality reduction should make sense and interpretation should be consistent and intuitive

Cardinality Reduction Examples

- Business intuition

Merchants	New Group
Walmart	SuperMarket
Target	SuperMarket
Macy's	Department Store
Dillards	Department Store

- Target based with cardinality reduction

	count	default	default_rate
purpose			
credit_card	525	207	0.394286
major_purchase	115	49	0.426087
home_improvement	186	80	0.430108
debt_consolidation	1738	862	0.495972
other	318	172	0.540881
medical	30	19	0.633333
small_business	88	56	0.636364

- WOE with cardinality reduction

Scaling as Transformation

In real data, it is normal two variables may have different scales, for example house age and price. But from the machine learning point of view, how these two columns can be compared, some ML algorithm required comparable feature(PCA...), sometime scale transformation can fix problematics data and can increase algorithm converge efficiency.

- ➔ Normalization(min-max scaling)

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- ➔ Standardization (Z-score standardization)

$$z = \frac{x - \mu}{\sigma}$$

Feature Creation and Data aggregation

Feature Engineering steps:

- Inspection existing data
- Brainstorm features.
- Create features.
- Check how the features work with the model.
- Start again from first until the features work perfectly.

Feature Creation Examples

This is a date column in a dataset, what will you do to use it create new variables?

date
2017-01-01
2008-12-04
1988-06-23
1999-08-25
1993-02-20

Year / Month / Week-of-day

Year since today(or another date)

Month since today(or another date)

Before or after crisis/add more enco data

	date	year	month	passed_years	passed_months	day_name
0	2017-01-01	2017	1	2	26	Sunday
1	2008-12-04	2008	12	11	123	Thursday
2	1988-06-23	1988	6	31	369	Thursday
3	1999-08-25	1999	8	20	235	Wednesday
4	1993-02-20	1993	2	26	313	Saturday

Feature Creation Examples

This is a flight time and status data? What is the new feature you can create?

	Date_Time_Combined	Status
0	2018-02-14 20:40	Delayed
1	2018-02-15 10:30	On Time
2	2018-02-14 07:40	On Time
3	2018-02-15 18:10	Delayed
4	2018-02-14 10:20	On Time

Flight Date Time Data

Hour of a day

Morning/noon/afternoon/evening/midnight

Feature Creation Examples

This is a credit card usage data. What is the new feature you can create?

Customer	Credit Limit	Card Spending
A	$5,000	$4,950
B	$8,000	$100
C	$20,000	$10,000
D	$2,000	$1,600

Utilization

Data aggregation Example

Transactional data, real time data handling require data aggregation and innovative feature engineering, here is an fraud example to use for discussion:

Merchant id	Merchant category code	Merchant city	Merchant state	Time	Transaction method	Transaction type	Amount
K2203	BC	BOS	MA	9:02	Magnetic	Retail	100.10
L3425	GD	NYC	NY	9:10	Magnetic	Retail	40.10
F3928	VS	NYC	NY	10:20	Chip	Retail	5.10
W9843	TY	POR	ME	13:20	Magnetic	Internet	200.00

Anything suspicious?

City	first_trans_time	last_trans_time	time_diff	distance_diff
BOS	8:01	9:02		
NYC	9:08	10:20	8	250

A fraudster will try to abuse the card as much as possible in a short period of time before the card is detected and suspended. So we should see abnormal transactions in a short period of time

source

Data aggregation Example

Aggregation of the transactions for each customer, either min, max, mean or sum, can reveal a lot insights with hundreds of variables:

➔ Aggregation by time:

Mean/Max amount spent per-transaction/per-day/by merchant in past 3, 5, 7, 2-weeks, n-weeks

➔ Aggregation by merchant category

Daily spending amount/recent transaction count/recent amount compare to past behavior for each merchant or group of similar merchant

➔ Aggregation by merchant location and time:

Number of retail location per day over past 1 week, 2 week, n-week

➔ Aggregation by transaction method:

Mean amount/number transaction spent by transaction method per-day in past 1 week, 2 week, nweek

source

Data aggregation Example

- ✗ The average amount spent per transaction over a month on all transactions
- ✗ The average amount spent over the course of 1 week during the past 3 months
- ✗ The average amount spent per day over the past 30 days
- ✗ Average amount per day spent over a 30 day period on all transactions up to this one on the same merchant type as this transaction
- ✗ Total number of transactions with the same merchant during the past 30 days
- ✗ The average amount spent over the course of 1 week during the past 3 months on the same merchant type as this transaction
- ✗ Total amount spent on the same day up to this transaction
- ✗ Total number of transactions on the same day up to this transaction
- ✗ Average amount per day spent over a 30 day period on all transactions up to this one on the same merchant as this transaction

source

Model Interpretability

Interpretable model

Regression, Naive Bayes, Decision Tree are interpretable models:

- Model equation/coefficient will help interpret the model
- Visualize the model could help for interpretation

What's the key points by interpreting the model:

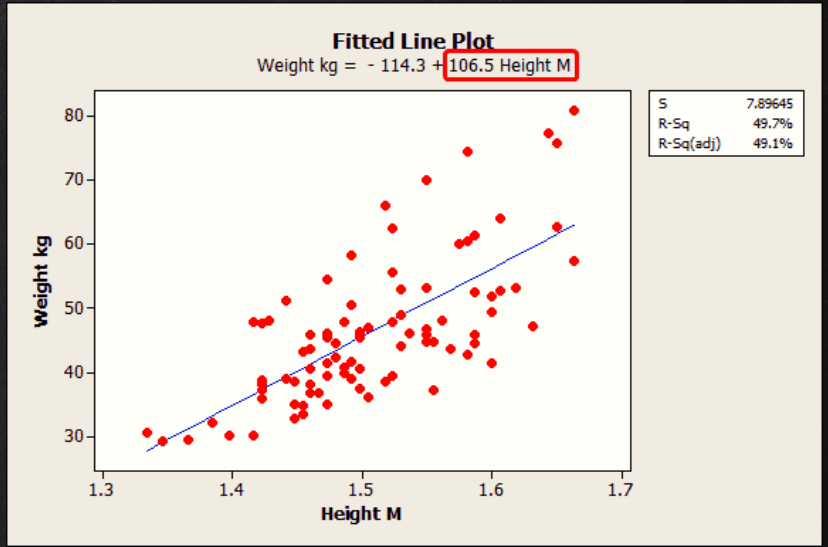
- Are all predictors have intuitive influence/relationship with target
- What are the key driver for your problem
- What's the impact sizing related to each key driver
- Any other business insights learnt from
- Any short term/long term benefit from the solution

Coefficient Analysis

- Linear regression coefficients

Coefficients

Term	Coef	SE Coef	T	P
Constant	-114.326	17.4425	-6.55444	0.000
Height M	106.505	11.5500	9.22117	0.000



- P value and coefficient for variables

Coefficients

Term	Coef	SE Coef	T	P
Constant	389.166	66.0937	5.8881	0.000
East	2.125	1.2145	1.7495	0.092
South	5.318	0.9629	5.5232	0.000
North	-24.132	1.8685	-12.9153	0.000

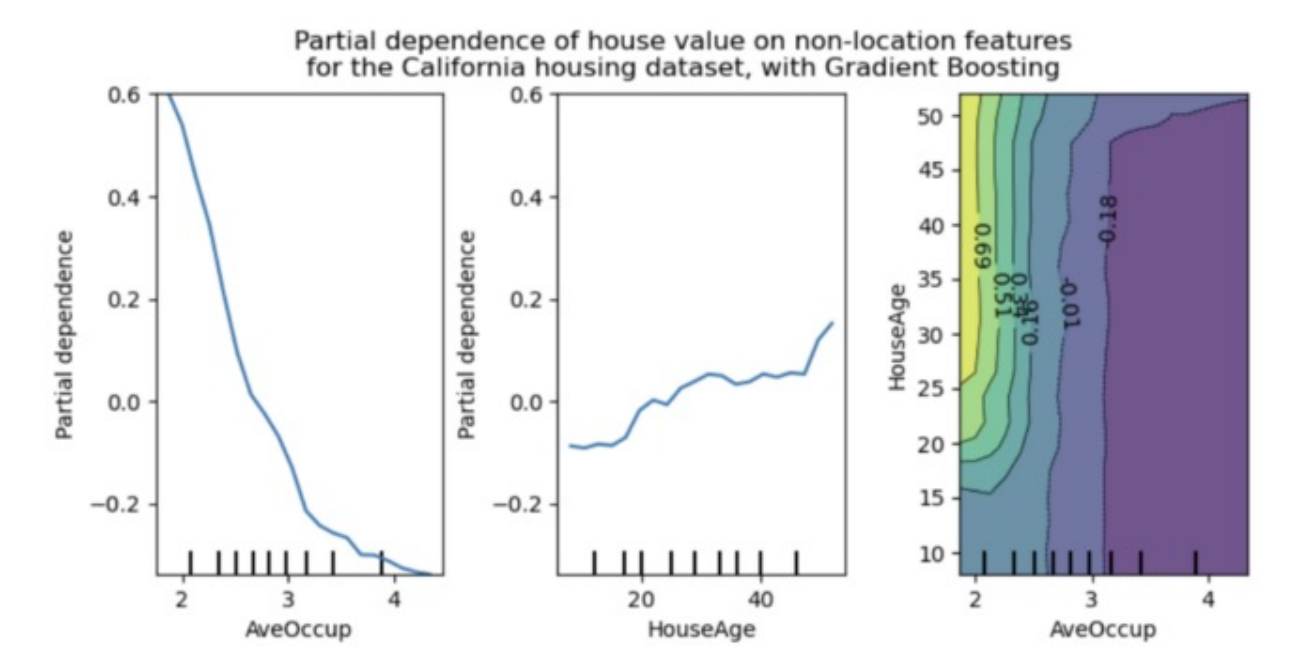
source

Methods to interpreting black-box models

1. Feature Importance
2. PDP (Partial Dependence Plot) & ICE
3. Shapley
4. LIME (Local Interpretable Model agnostic Explanations)
5. Prediction Decomposition

Partial Dependence

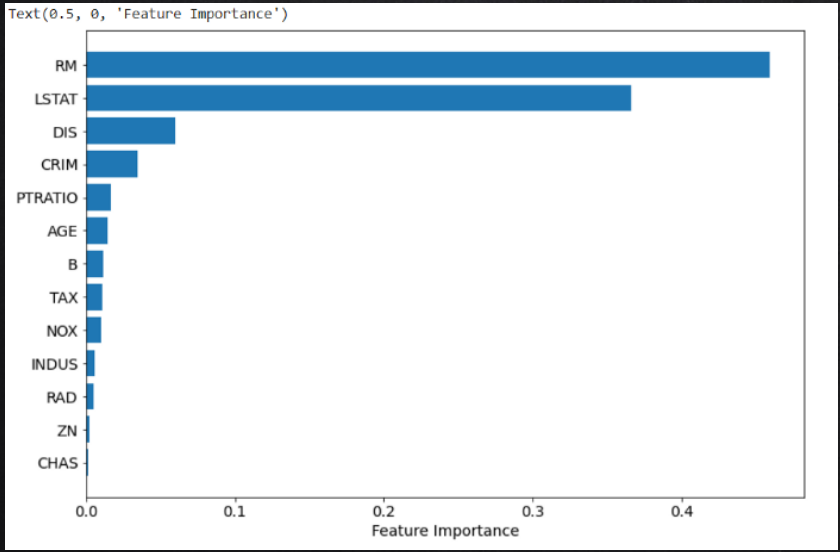
✗ Can be applied to tree-based models



source

Feature Importance

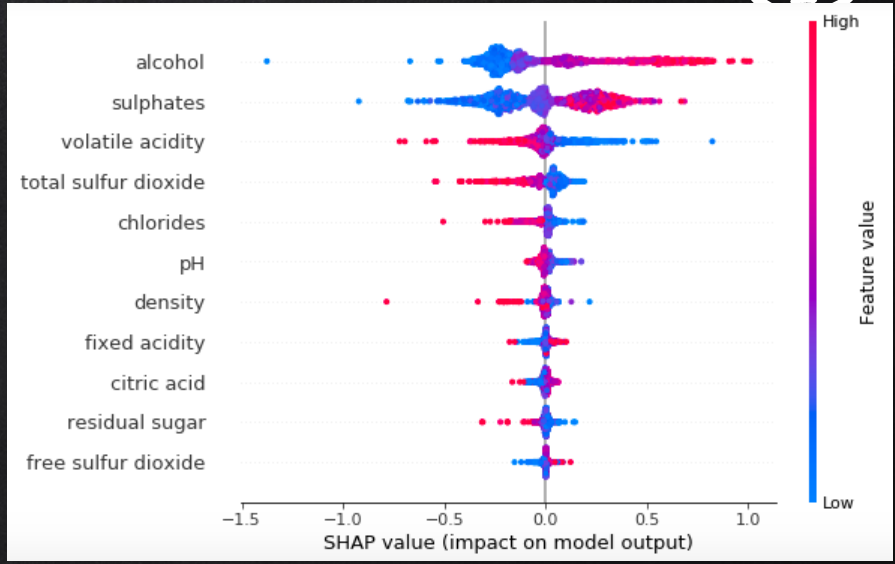
- ✗ Gini index
- ✗ Permutation Feature Importance

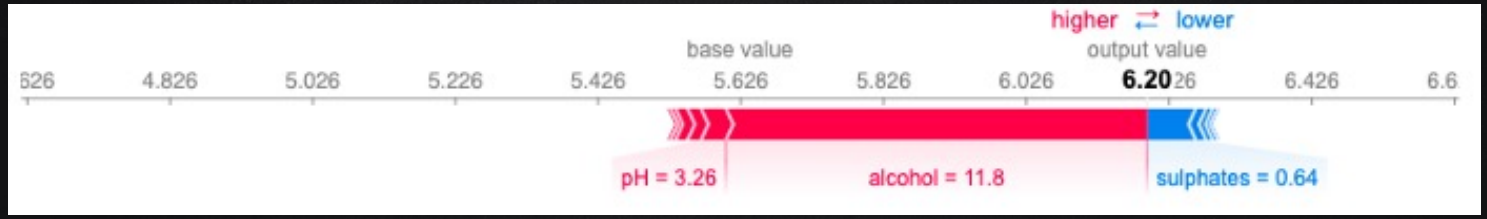


source

Feature Importance

- ✗ SHAP plot to show relationships of predictors and target variable
- ✗ Individual SHAP Value plot for local interpretability





source

Recommended Readings

https://www.analyticsvidhya.com/blog/2018/03/introduction-regression-splines-python-codes/

https://statisticsbyjim.com/regression/interaction-effects/#:~:text=For%20example%2C%20a%20three%2Dway,on%20both%20Food%20and%20X.

https://towardsdatascience.com/interpretability-in-machine-learning-70c30694a05f

https://towardsdatascience.com/interpreting-machine-learning-model-70fa49d20af1

https://www.kdnuggets.com/2017/11/interpreting-machine-learning-models-overview.html

https://lilianweng.github.io/lil-log/2017/08/01/how-to-explain-the-prediction-of-a-machine-

https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/

References

Bruce, Peter,Bruce, Andrew,Gedeck, Peter. Practical Statistics for Data Scientists (Kindle Location 5267). O'Reilly Media. Kindle Edition.

https://www.theanalysisfactor.com/seven-ways-to-make-up-data-common-methods-to-imputing-missing-data/

https://towardsdatascience.com/a-data-scientists-toolkit-to-encode-categorical-variables-to-numeric-d17ad9fae03f

https://towardsdatascience.com/a-data-scientists-toolkit-to-encode-categorical-variables-to-numeric-d17ad9fae03f

https://towardsdatascience.com/how-to-create-good-features-in-fraud-detection-de6562f249ef

learning-model.html

https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/

https://scikit-learn.org/0.22/auto_examples/inspection/plot_partial_dependence.html

https://towardsdatascience.com/understanding-feature-importance-and-how-to-implement-it-in-python-ff0287b20285

https://medium.com/dataman-in-ai/explain-your-model-with-the-shap-values-bc36aac4de3d