Iron Hack Data Analytics Course Notes

P. Zimmerman

November 9, 2022

1 Acronyms

- EDA - Exploratory Data Analysis. (Interact with through visualization of stats)
- ETL - Extract, Transform, and Load
- CRUDE - create, update, extract?
- MAE - mean absolute error

2 General DA Terminology

- Descriptive vs Predictive analytics
- **supervised learning algorithms** Algorithms which experience a dataset containing features, but each example is also associated with a label or target.

3 Key Performance Indicators (KPIs)

- • are performance markers forming the core of every business. KPIs serve to provide information about where to dedicate resources, focus sales, optimise costs, and other underlying overarching objectives.

  • Profit = Revenue - Expenses

  • Gross Profit = Revenue - Direct Costs
  (related to what's being sold, server costs, essential costs per product item)

- Net Profit = Gross Profit - Indirect Costs
(salaries, rent, etc)

- Retention Rate = 1 - Churn Rate

- Outcomes

- Conversion Rate - How many visitors/users become customers (marketing funnel)

4 Data Wrangling

- rename columns in lower case to be more descriptive (using functions). Option inplace=True reassigns the variable when renaming.

- Cleaning: df.info() check

1. Look for format. Are the quantities “objects”?

2. Average values for large numbers (income) are often ints 3. Average values for lower numbers ( ) could be floats

5 Health care for all study

-

- Assumptions: people who have donated in the past will donate again and those who lapsed are are unlikely do donate again. Linear regression criteria.

- target variable : `target_d` is total amount donated

- Objective : Profit, given costs like mailing asking for donations. Use machine learning (linear regression) to produce a ranking of donors by how likely they are to respond and how much they are likely to donate. .

- important input variables
- Observation: Donors who donated large amounts are less likely to respond to the mail. Inverse correlation between amount and response.

- From the sample, we want to build a model which predicts who is going donate the most amount of money. Only including those who are most likely to respond is biased to low dollar amount donors.

- Assumption is that the highest value donors have the largest average donations.

6 Car Insurance Policy study

- Assumptions: Linear regression criteria.

••••••

- Target variable : `total_claim_amount`

- Observations: See plots from 21/06/2022

- Objective : Find out what's influencing policy costs using machine learning (regression) to produce a model predicting customer premiums from variables like policy_type and vehicle_size.

- Important features

- `monthly_premium_auto` is input feature most correlated with the target variable, carrying a Pearson coefficient of $r = 0.63$.
- In the model, where we have used encoded categorical variables, we find that the most important features are
  1. `encode_location_code_suburban`
  2. `encode_location_code_urban`
  3. `encode_location_code_rural`
  4. `encode_coverage_basic`
  5. `encode_coverage_extended`

7 Modeling

- classification - for categorical (discrete) variables. Predict letters, animal type, facial recognition for race/gender.
- **linear regression** - . Extrapolations based on fitting data

- Input: features/ explanatory variables/predictor variables/independent variables
- Output: labels/target/explained variable/predicted variable/dependent variable/y
- Linear Regression Assumptions

  1. Linearity
  2. Independency
  3. Normalcy of residual errors
  4. Homoscedasticity

- regressor - variables in a regression model, excluding the constant
- Standardization : standardization, where the fields are transformed to have zero mean and unit variance, is a common preprocessing routine useful (or even necesssary) for optimal predictability of several many machine learning estimator algorithms (such as the RBF kernel of Support Vector Machines or the L1 and L2 regularizers of linear models) implemented which are implemented scikit-learn.
- test/train split
- A design matrix, usually denoted X, is a table containing a different observation in each row. Each column of the matrix corresponds to a different feature.
- Zvariable $z = (x - \mu)/\sigma$ and its lookup table.
- 95% confidence interval
- Z-test and two-tail test (2.5%: $z = 1.96$) one-tail test (5%: $z = 1.65$)
- degrees of freedom dof $= n - 1$, where n is the number of examples (rows).
- **mean squared error**

  $$\text{MSE}_{\text{test}} = \frac{1}{m} \sum (\hat{\vec{y}}^{(\text{test})} - \vec{y}^{(\text{test})})_i^2, \tag{1}$$

  where $\hat{y}$ is the model outcome. If y is new data not found in the training set, then the MSE refers to the so-called *generalization error*.

- The coefficient of determination (R^2 measure),

  $$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}, \tag{2}$$

where y is the true value, $\hat{y}$ is the predicted target variable, $\bar{y}$ is the sample mean, and the sum is over all the samples. For data centered around the mean $\bar{y} = 0$,

$$R^2 = 1 - \frac{\mathrm{Var}(y - \hat{y})}{\mathrm{V}ar(y)}, \qquad \bar{y} = 0. \tag{3}$$

The value of R^2 increases with the number of variables, as it assumes that each variable is influencing the prediction of the target variable regardless of feature importance (really?) . A related quantity

$$R^2_{adj} = 1 - \frac{(n-1)(1-R^2)}{(n-1-k)}$$

is more robust to overfitting, as it only assumes that the model is affected by important features. The adjusted R^2 will penalise you for adding independent variables (via k in the equation) that do not fit the model. Why? In regression analysis, it can be tempting to add more variables to the data as you think of them. Some of those variables will be significant, but you can't be sure that significance is just by chance. The adjusted R2 will compensate for this by that penalizing you for those extra variables.

- **regularization** - Regularization is any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error.
- **validation set** - A set of samples, which the training algorithm does not observe, used to estimate generalization error and guide selection of hyperparameters.
- **cross validation** - It is used to estimate generalization error of a learning algorithm when the given dataset is too small for a simple train/test or train/valid split to yield accurate estimation of generalization error, because the mean of a loss on a small test set may have too high a variance.
- **Point estimator or statistic** A general function of the data $\hat{\vec{\theta}} = g(\vec{x}_1, \ldots, \vec{x}_n)$.
- **bias**

  $$\mathrm{bias}(\hat{\vec{\theta}}) = \mathbb{E}(\hat{\vec{\theta}}) - \vec{\theta}. \tag{4}$$

  The bias of an estimator is related to the variance by

  $$\begin{aligned} \mathrm{MSE} &= \mathbb{E}[(\hat{\vec{\theta}} - \vec{\theta})^2], && (5) \\ &= \mathrm{bias}^2(\hat{\vec{\theta}}) - \mathrm{Var}(\vec{\theta}). && (6) \end{aligned}$$

  There is a tradeoff between bias and variance related to under of overfitting.

- **likelihood function** - joint probability of the observed data viewed as a function of the parameters of the chosen statistical model, $\mathcal{L}(\vec{\theta}|\vec{x}) = p_{\text{model}}(\vec{x};\vec{\theta})$. It is viewed as a function of parameters $\vec{\theta}$ with the observations fixed.
- maximum likelihood estimator

8 Relational Databases (RDMS)

Structured Query Langauge Postgre, MySQL, BigQuery, Amazon REDSHIFT Structured data (e.g. columns and rows) stored in several related tables (pandas DFs). Fixed scheme ACID compliancy (atomic,consistent,isolated,durable) normalized data 1. In SQL the "schema" is the database 2. CRUD operations 3. data types: binary=boolean 4. data encoding (unicode character: utf-8 is most popular and can support chinese and arabic) 5. SQL tables are saved in the disk, in comparison with pandas DFs which are saved in RAM 6. Use SQL dump to recreate table. Server/Data Export