DATA8001 - Assignment 1

Student ID: R00207204

Student Name: Bernard McNamee

Report

Summary

What is the question? *What is the best linear regression machine learning model to predict car prices using a dataset of car sales from previous years?*

Tasks

1. Dataset imported into python as a Pandas dataframe and cleaned using python ETL functions, eg replicating data values where missing from other columns and standardising values by format.

2. Five specific report questions were answered with visualisations and numeric answers.

3. An iterative modeling process created 4 models with own python functions and sklearn machine learning module functions starting with all features and gradually reducing to three features that generated the best Rsq score:

    1. features were selected

    2. dataset was split 80:20 into training and test dataframes

    3. label encoding was applied to categorical text variables (features)

    4. scaler encoding was applied to all numeric variables (features)

    5. a linear regression model was created with training data

    6. model was tested against test data → RMSE / RSq accuracy scores

    7. accuracy scores were calculated for individual features

4. Results were validated using own R test functions

    1. checking for Rsq score for each model

    2. checking correlation between variables, multicollinearity and interaction between variables, to find clues to the best model.

Results

1. Model 1: all features, label encoding - RMSE: 21416 RSq: 0.235
2. Model 2: all features, label/scaler encoding - RMSE: 21416 RSq: 0.235
3. Model 3: 'make','model','tax_band' features, label encoding - RMSE: 21515 RSq: 0.228
4. Model 4: 'make','model','tax_band','make_model_tax_band' features, label encoding - RMSE: 20869 RSq: 0.274 **(ACCEPTED)**
5. Model 5: On exploring the dataset, it was found to be right skewed. A normalised dataset (outliers removed) input to Model 4 generated much improved scores, RMSE: 10904 RSq: 0.489, when tested against its own test data. However, this dropped when tested against unseen data to RMSE: 21361 RSq: 0.239

Conclusion

The final model accuracy Rsq score 27% is poor – this is not a good model. Why is it not a good model? Perhaps the data is low quality, eg the car prices are not consistent and vary wildly.

Concern for the correct application of tools and techniques was addressed by replicating the process in R – the low score results matched closely. Also, removing outliers generated worse results – normalising data removed part of the underlying pattern.

The final model selected has just four features – 'make', 'model', 'tax_band', and an artificial feature, 'make_model_tax_band' features, combining the other features. The additional 4% gained using the artificial feature is a close call – considering parsimony, it might be better to just use the first three or just the artificial feature both at Rsq 24%.

The project was both frustrating and rewarding. I was confident coding in python but learning dataframes was difficult and time consuming. Also, while the modeling process was explained clearly in lectures, I still found it very hard to abstract from the lecture single feature examples to multiple feature modeling. In both cases, this was less to do with modeling experience and more to do with python experience.

Although the final model is disappointing as a predictive model, the project was a success when I consider that I learned a great deal wrt the mechanics of solving a machine learning problem. What I most enjoyed was how this project brought in concepts and methods from other modules such as Statistics last term and Maths & Modeling this term.

Question 1 - Discuss your understanding of companies that are considered to be 'analytic competitors'.

(max 300 words)

Thomas Davenport *(Competing On Analytics)*, defines an *analytical competitor* as an organisation that uses analytics extensively and systematically to out-think and out-execute their competitors. He further defines an *analytical competitor* as the top tier of a 'food pyramid' with successive tiers below having decreasing analytical capacity to the bottom tier with negligible capacity (running blind).

The *analytical competitor* has four key characteristics; 1) analytics supports a strategic and distinctive capacity – the company has identified its distinctive capacity, ie what sets it apart from competitors, eg revenue management, Marriott or customer service and loyalty, Caesers, and aggressively exploits analytics to boost this capacity; 2) an enterprise analytics approach/management – analytics is managed centrally and distributed broadly throughout the company; 3) senior management commitment – board level champion with passion for analytics, eg Jeff Bezos at Amazon, to drive change in culture, behaviour, processes, skills and technology; 4) great ambition wrt analytics – makes a strategic bet that analytics will drive growth and profitability when this change appears radical but in hindsight will seem logical and rational, eg Capital One, a credit card company with big ambitions, went beyond the industry norm of evaluating/pricing new business based on FICO credit scores, and designed a new business channel with partners that uses better analytics to determine which customers might be lower risks than their FICO score might predict.

Furthermore, the *analytical competitor* organisation is systematic and may use a variation of the DELTTA model to apply an analytics strategy: D – Data, E – Enterprise, L – Leadership, T – Targets, T – Technology and A – Analysts. Data, Enterprise and Leadership are as described above. Targets – applies analytic focus to key area, ie distinctive capacity, not all areas of business, Technology – new and complex big data technology investment, and Analysts – analyst team expertise investment.

Finally and most importantly, for this student of analytics, I believe while it's a cliché that any initiative needs CEO backing to succeed, this is an imperative and companies that have CEOs with analytical experience or training have a greater chance of succeeding, eg Reed Hastings at Netflix who have produced super successful tv shows inhouse, such as *House of Cards*, informed viewer preference analytics – now that is a major departure from mainstream programming thinking.

Question 2 - What is the difference between data cleansing and data validation, provide examples to demonstrate your answer?

(max 300 words)

Data cleaning differs from data validation in that validation seeks to check quality of data, whereas, data cleaning seeks to fix or remove anomolies. Typically data validation is the first step before data cleaning and serves as a gate to data cleaning. Validated and cleaned data will eventually reside in a database application to serve user queries or reports so the user experience will ultimately reflect the quality of the data.

Data validation is a method for checking data for

- validity – conforms to business rules/constraints, eg data type constraints, range constraints, regular expression patterns
- accuracy – degree to which data is close to true values, though valid values may not be true, eg a valid street address mightn't actually exist
- completeness – all required data is known
- consistency – reducing inconsistent contradictions
- uniformity – eg matching currencies, units of measure

Data Validation can be done manually (slow and expensive) or through automation with off the shelf tools or custom made scripts (fast, less expensive in long run).

Data cleaning removes or modifies data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted, for example:

- type conversion – change numerical data stored as strings to floating point
- typographical - syntax/spelling errors

Both are data processing processes/methods to increase quality of data, though the definition of constituent parts varies/overlaps, they are both data cleaning generally speaking. In a nutshell validation is an evaluation *(what data to use?)* and cleaning is a transformation *(improve data quality in preparation for analysis).*