Predicting & understanding bookmaker choice using machine learning methods

University of Hertfordshire UH

Katharine Chorlton

Supervisor: Dr. Carolyn Devereux

Background:

- Online sports betting community website www.olbg.com (OLBG) provides visitors with betting tips, offers of free bets and access to the best bookmaker deals.
- OLBG have supplied 12 months of their click-through data showing the specific bookmaker a visitor clicked through to. Data is anonymised so does not require ethical approval.
- The data includes over 1.2m records with each record representing one click through to a particular bookmaker. The data has various features which represent the who, what, where, and when of the click. (See Figure 1). All data is categorical and many variables have high cardinality (See Figure 1).

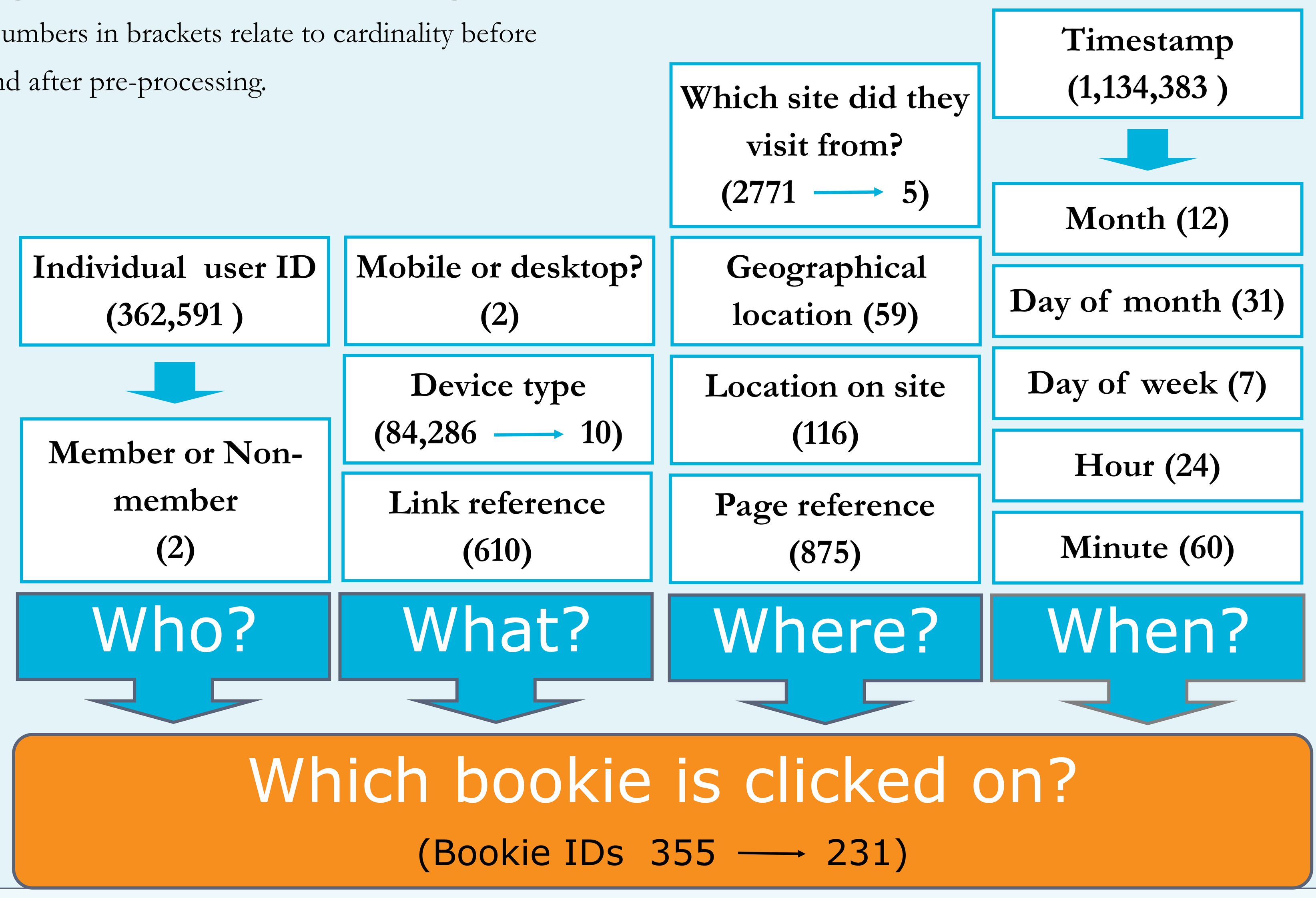
Aim: To compare the ability of several machine learning techniques to predict which bookmaker will be clicked on and the key factors influencing this.

Objectives

- **To interrogate the data and ascertain optimal machine learning algorithms**
- **To achieve a high score on a suitable prediction metric**
- **To identify key features**

Figure 1: Data variables that might influence which bookie is clicked on.

Numbers in brackets relate to cardinality before and after pre-processing.



Literature Review:

Studies of online betting did not prove relevant. Further work focused on identifying studies with comparable data and aims.:

Multiclass classification of categorical variables

Who? Gupta et al., 2017

What? Decision trees are an effective technique to handle categorical data.

Relevance: Decision trees can be visualized so are easily interpretable; which is important to OLBG.

Analysis: The study provides a clear and concise critique of different decision tree methods but neglects to mention more advanced algorithms such as gradient boosted decision trees.

Potdar et al., 2017: Most machine learning methods require categorical variables to be converted to numeric data.

Who? Pargent et al., 2022

What? A large scale study looking at the different methods of encoding categorical variables.

Relevance: Particular focus on high cardinality variables.

Analysis: Dummy encoding is unsuitable for high cardinality variables. Target encoding can produce superior results, but run times are slower.

Target encoding increases the risk of overfitting on the training data

Who? De La Bourdonnaye & Daniel, 2021

What? Aim to conclude whether encoding categorical data via preprocessing is superior to built-in encoding. Analysis was conducted on a credit card fraud detection database.

They compare 3 gradient boosted decision tree algorithms:

Extreme Gradient Boosting (XGBoost) Needs encoded data	**Light Gradient Boosted Machine (LightGBM)** Built-in encoder	**CatBoost** Built-in encoder

Relevance: Data had high cardinality categorical variables.

Analysis: The CatBoost algorithm produced the best results and was the most easily implemented.

Method:

Feature engineering was used to reduce the cardinality of some variables and the target (see Figure 1).

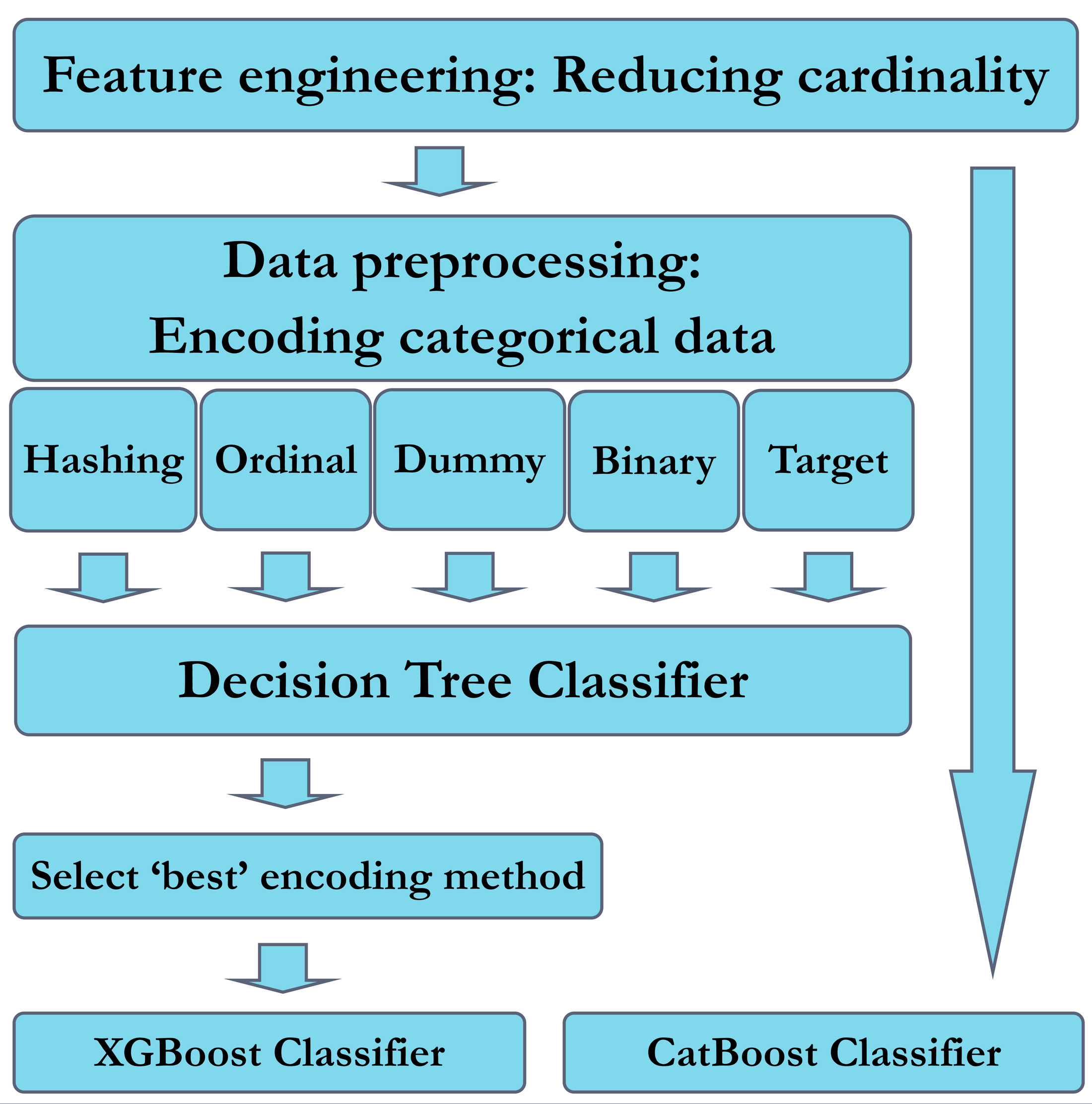
Key encoding methods to test were identified (see Table 1).

Table 1: Description of encoding methods to test

Encoding method	Description
Hashing	Turns arbitrary features into indices in a vector or matrix by applying a hash function (used in data encryption).
Ordinal	Each unique category value is assigned an integer value.
One-hot/ dummy	Transforms all the elements of a categorical variable into new columns represented by 0 or 1 (binary values) to signify the presence of the category value.
Binary	Categorical feature converted into numerical using ordinal encoder. Then transformed to a binary number and split into different columns.
Target	Replace a categorical feature with average target value of all data points belonging to the category.

These encoding methods were tested on a random sample of the data (10%) using a basic Decision Tree Classifier selected for its ease and speed of implementation (See Figure 2).

Figure 2: Diagram showing process undertaken



Preliminary Results:

Algorithm: Decision Tree Classifier.

Data: A sample of 10% of total data. Split 70% train, 30% test.

Metrics: Accuracy, area under the ROC curve (AUC).

Aim: Select optimal encoding method.

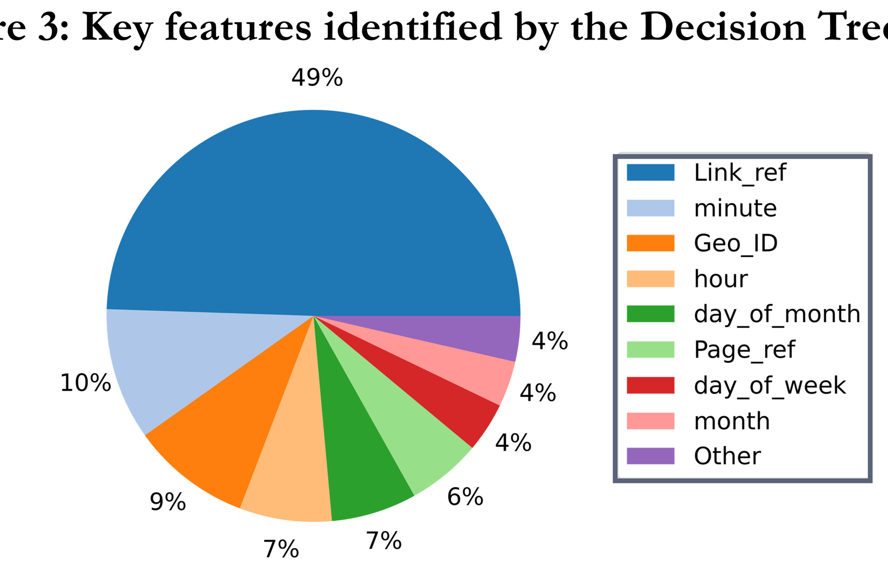
Table 2: Results of different encoding methods

Encoding method	Accuracy	AUC
Hashing	62.73	0.75
Ordinal	67.88	**0.84**
Dummy	62.96	0.76
Binary	62.13	0.74
Target	**67.98**	0.82

Result: Ordinal encoding selected as the optimal method (see Table 2). Target encoding discounted due to risk of overfitting.

The Decision Tree Classifier showed the most important feature was link ref (see Figure 3).

Figure 3: Key features identified by the Decision Tree



Further Results:

Algorithm: XGBoost Classifier vs. CatBoost Classifier

Data & Metrics: as above

Table 3: Preliminary results from advanced algorithms

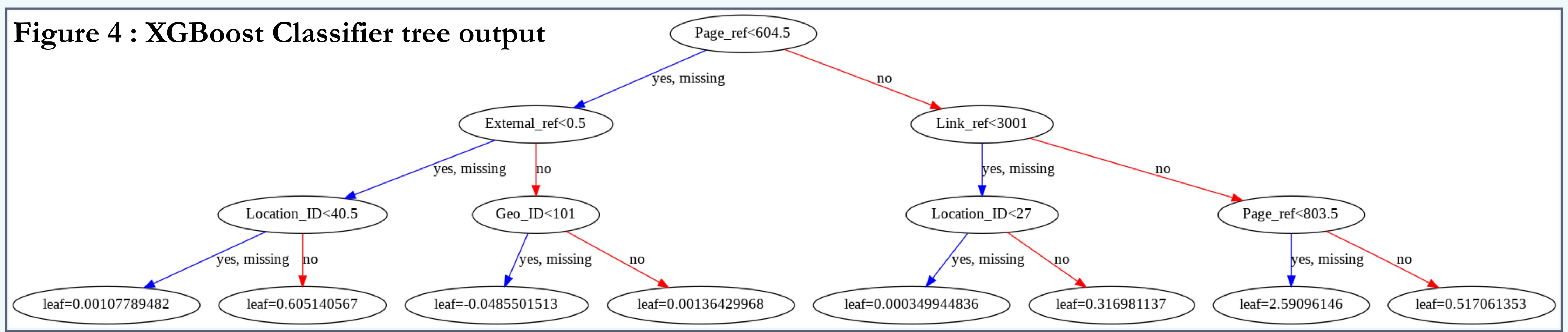
Classifier	Accuracy	AUC	Notes
XGBoost	72.47	0.87	Higher than Decision Tree
CatBoost	27.29	0.64	Much lower - needs further work

Result: Table 3 shows very poor results for the CatBoost algorithm which is surprising given the literature reviewed. Figure 4 shows the tree diagram for the XGBoost model. The key feature used to split the tree is page ref; different from the Decision Tree's key feature (see Figure 3).

What's next?

- Continue to investigate the most appropriate metric on which to 'score' results
- Optimize parameters for XGBoost classifier and CatBoost classifier
- Run models on full data set
- Draw conclusions: which model is the most appropriate for this data

Figure 4 : XGBoost Classifier tree output



References

Gupta, B., Rawat, A., Jain, A., Arora, A. and Dhami, N., 2017. Analysis of various decision tree algorithms for classification in data mining. International Journal of Computer Applications, 163(8), pp.15-19.

De La Bourdonnaye, F. and Daniel, F., 2021. Evaluating categorical encoding methods on a real credit card fraud detection database. arXiv preprint arXiv:2112.12024.

Pargent, F., Pfisterer, F., Thomas, J. and Bischl, B., 2022. Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. Computational Statistics, pp.1-22.

Potdar, K., Pardawala, T.S. and Pai, C.D., 2017. A comparative study of categorical variable encoding techniques for neural network classifiers. International Journal of Computer Applications, 175(4), pp.7-9.