11-2019

Using Machine Learning to Predict Sales Conditional on Bid Acceptance

Barry E. King

Jason Davidson

Using Machine Learning to Predict Sales Conditional on Bid Acceptance

Barry E. King, Jason Davidson

Butler University – Lacy School of Business, 4600 Sunset Avenue, Indianapolis, IN 46208, United States

Abstract

A North American provider of vehicle parking solutions seeks to predict if a bid will be successful and, for those that are successful, what will be the cumulative sales revenue. Both traditional statistical methods and machine learning algorithms were employed. The machine learning techniques performed better than the statistical methods. There is no statistically significant difference between random forest and extreme gradient boosting for either the binary classification task or the regression task.

Keywords – *logistic regression, linear regression, random forest, extreme gradient boosting, Tukey honestly significant test*

I. INTRODUCTION

Predicting sales conditional on winning a bid is a two-fold prediction problem. First, given a variety of predictor variables and a history of winning or losing bid sales, will the sales bid be successful or unsuccessful? If the bid is successful, what will be the cumulative revenue from the sale? Machine learning methods are employed in addition to traditional statistical methods. The machine learning approaches outperform the traditional methods for both forecasting tasks.

II. THE PROBLEM

A North American provider of parking technology solutioins wishes to predict if a production adoption bid will be successful. The company would like to determine what predictor variables influence customer adoption. Furthermore, can cumulative revenue be predicted?

III. DATA

The company has recorded nearly 28,000 observations from sales prostpects spanning 2006 through 2019, of which 1440 are lost bids. It is important to note, some sales data were recorded in Canadian currency. These values were converted to United States dollars for this study.

A. Skewed Sales

Due to small lower boundaries that are often associated with financial data, sales data are skewed-right as evidenced in Figure 1.

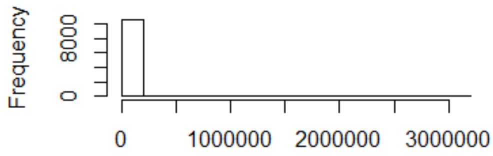

Fig 1: Skewed sales data

A log transform was applied to the sales data, a seen in Figure 2. The log transformation allows for clear interpretation of data against the original scale.

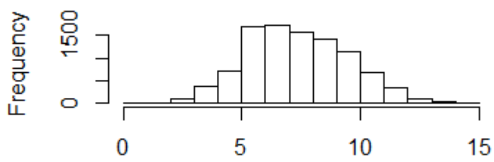

Fig 2: Distribution of log of sales

Log of sales replaced sales as the target variable for the conditional sale prediction task.

IV. FEATURES

The raw data contained fourteen variables. Most, like customer ID or opportunity were unusable for analysis.

A. Population and Per Capita Income by State

A state's population and per capita income were obtained from the US Census (2019). These numeric variables were merged into the data on the state where the sale was made.

B. Create Dummy Variables

The company sells seven types of products in seventeen states. The state and product variables were made into dummy variables using caret's dummyVars function.

C. Feature Reduction

Feature reduction was performed using the Boruta feature selection method rather than Akaike Information Criterion. Boruta is a tree-based method.

D. Collinearity

Boruta does not check for collinearity. Variance inflation factor was applied to increase the stability of the regression and reduce the standard error by decreasing the feature set further.

E. *Final Feature Set*

The final feature set is reported in Table I.

TABLE I
Final Feature Set

Feature	Comment
Log of sales revenue	Target variable
Age	Days between date created and date closed
Date created	Date closed was dropped since it would be collinear with Age
State	17 possible states
Type	7 possible product types
Canadian	Binary variable. Were original sales dollars Canadian?
State population	Merged from US Census data
State per capita income	Merged from US Census data

V. METHODS FOR SUCCESSFUL OR UNSUCCESSFUL BID

The first prediction task was to classify an observation as a successful or unsuccessful bid. Stratified sampling was employed due to the low number of unsuccessful bids.

A. *Binary Classifiers*

Three binary classifiers were tuned and used on the historical data.

1) *Logistic Regression:*

Logistic regression is the traditional statistical method for predicting a binary classification.

2) *Random Forest:*

Random forest was chosen due to its robustness and success in other of the author's investigations.

3) *Extreme Gradient Boosting:*

Extreme gradient boosting was selected due to its considerable success in machine learning competitions such as the Kaggle competitions [1].

B. *Misclassification Rate*

Extreme gradient boosting was assessed to be the best method for the binary classification task with an out-of-sample misclassification rate of 4.0 percent. See Table 2.

TABLE II
In- and Out-of-Sample Misclassification Rates, Three Methods

Method	In-Sample Misclassification Rate	Cross-Validated Out-of-Sample Misclassification Rate
Logistic regression	0.049	0.048
Random forest	0.025	0.041
Extreme gradient boosting	0.036	0.040

1) *Boxplots of Cross-Validated Out-of-Sample Misclassification Rates:*

Figure 3 displays boxplots of the misclassification rate for three methods. It appears that logistic regression does not perform as well as the other two techniques, however random forest and extreme gradient boosting perform about as well.

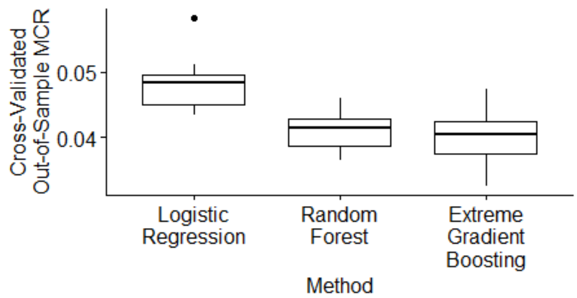

Fig 3: Boxplots of 10-fold cross-validated misclassification rates

2) *Tukey Honestly Significance Difference Test:*

Table 3 reports the significant differences between method pairs. Logistic regression performs differently than the other two methods but there is no statistically significant difference between the random forest and extreme gradient boosting.

TABLE III
Results of Tukey Honestly Significance Difference Test

Method Pairs	Difference	Lower	Upper	p Adjusted
Random Forest-Logistic Regression	-0.007	-0.012	-0.003	0.001
Extreme Gradient Boosting-Logistic Regression	-0.008	-0.012	-0.003	0.001
Extreme Gradient Boosting-Random Forest	-0.001	-0.005	0.004	0.936

C. *Regressors*

1) *Linear Regression:*

The data for linear regression were scaled to avoid the well-known problem of using unscaled data with linear regression. Large-valued features can dominate small-valued features.

2) *Random Forest:*

The mtry parameter of the randomForest function was optimized at 9.

3) *Extreme Gradient Boosting:*

Grid search was used on some of the extreme gradient boosting parameters to optimally tune the algorithm.

Figure 4 shows the relative importance of features to developing an accurate log sales forecast. Age and Date created dominate the importance. The merged variables, state per capita income and state population, appear in the top six features although they are relatively unimportant.

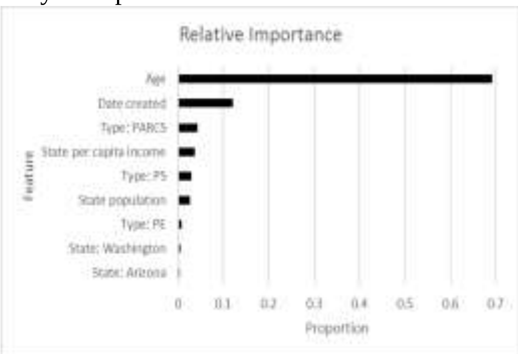

Fig 4: Relative importance of features when making cumulative sales predictions

D. *Root Mean Square Error*

Table 4 reports in-sample and out-of-sample root mean square error (RMSE) for the three algorithms being assessed. Random forest has the best out-of-sample RMSE.

TABLE IV
In- and Out-of-Sample Root Mean Square Error

Method	In-Sample RMSE	Cross-Validated Out-of-Sample RMSE
Linear model	1.967	1.974
Random forest	1.317	1.652
Extreme gradient boosting	1.238	1.675

1) *Boxplots of Cross-Validated Out-of-Sample RMSE*

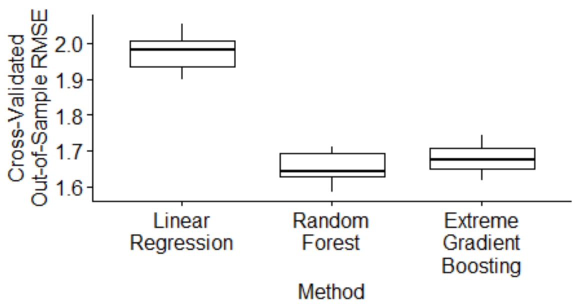

Fig 5: Boxplots of 10-fold cross-validated RMSE for three methods

2) *Tukey Honestly Significance Difference Test*

TABLE V
Tukey Honestly Significance Difference Test

Method Pairs	Difference	Lower	Upper	p Adjusted
Random Forest-Linear Regression	-0.322	-0.372	-0.272	0.000
Extreme Gradient Boosting-Linear Regression	-0.298	-0.348	-0.248	0.000
Extreme Gradient Boosting-Random Forest	0.023	-0.027	0.073	0.489

As with the binary classification task, there is no statistically significant difference between random forest and extreme gradient boosting with respect to performing the log sales forecast of a successful bid.

VI. CONCLUSIONS

Machine learning methods performed better than statistical techniques on this problem. Analysts are cautioned not to assume machine learning will always perform better than traditional statistical methods but should assess the performance of each on cross-validated out-of-sample analyses. Random forest and extreme gradient boosting performed about as well for both predictive tasks – binary classification followed by sales regression

REFERENCES

[1] G. Janson, What Machine Learning Approaches Have Won Most Kaggle Competitions?, Retrieved from https://www.quora.com/What-machine-learning-approaches-have-won-most-Kaggle-competitions.