Online Shoppers Purchasing Intention

Is it possible to predict whether a customer will buy a product from your website?

Agenda

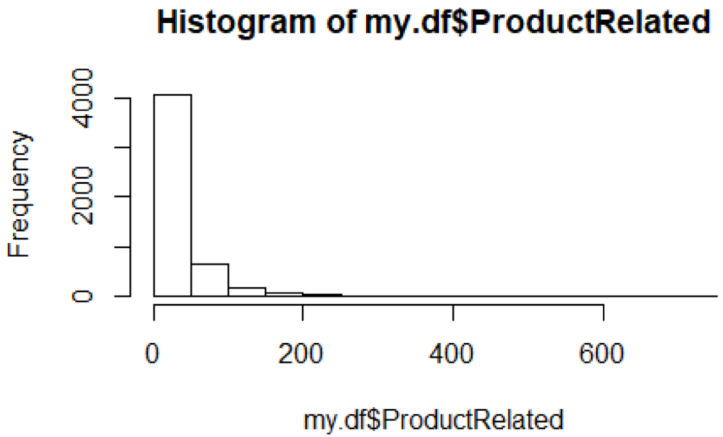
- Define Business Problem
- Data Cleaning
- Machine Learning Models
  - Principal Component Analysis
  - Lasso
  - Classification Tree
  - Support Vector Machine
  - Random Forest
  - Boosting
- Business Analytics

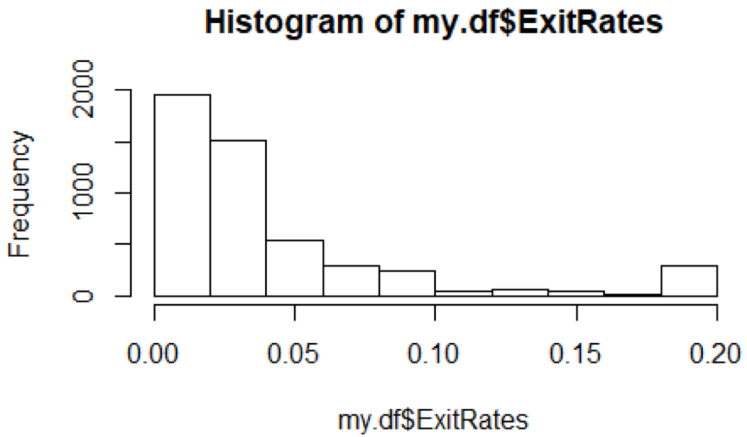
Dataset

- Source
  - UCI Machine Learning Repository
- Features
  - The dataset consists of feature vectors belonging to 12,330 sessions, 10 numerical and 8 categorical attributes.
  - The 'Revenue' attribute can be used as the class label.

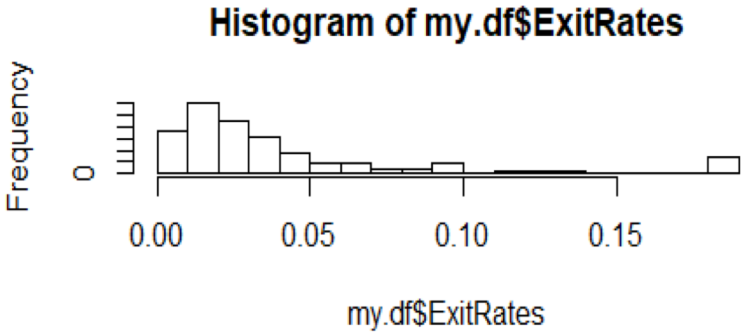
Histogram of Numeric Variables

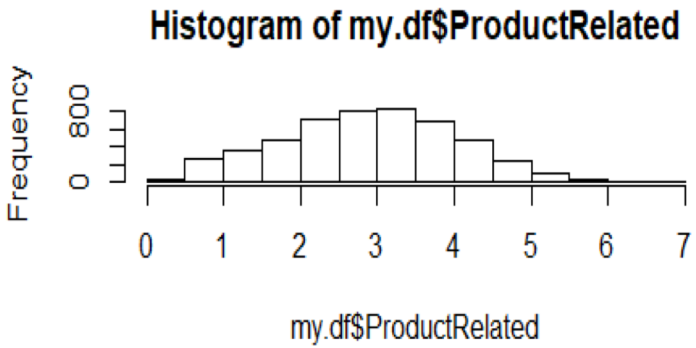
Right-skewed



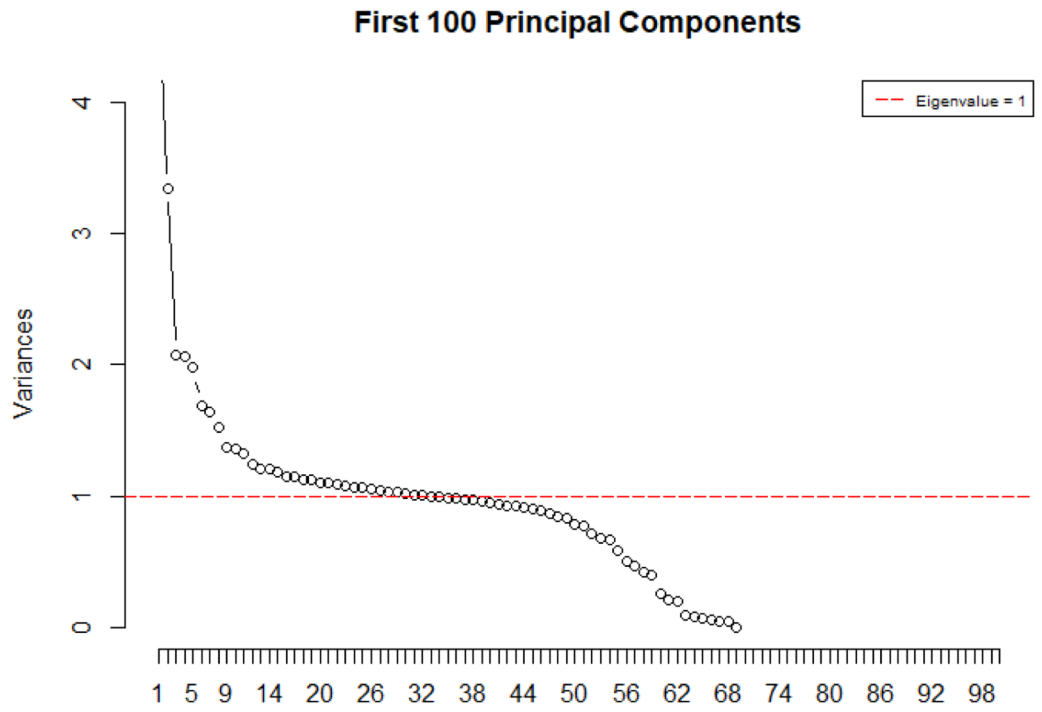


Log Transformation





Principal Component Analysis

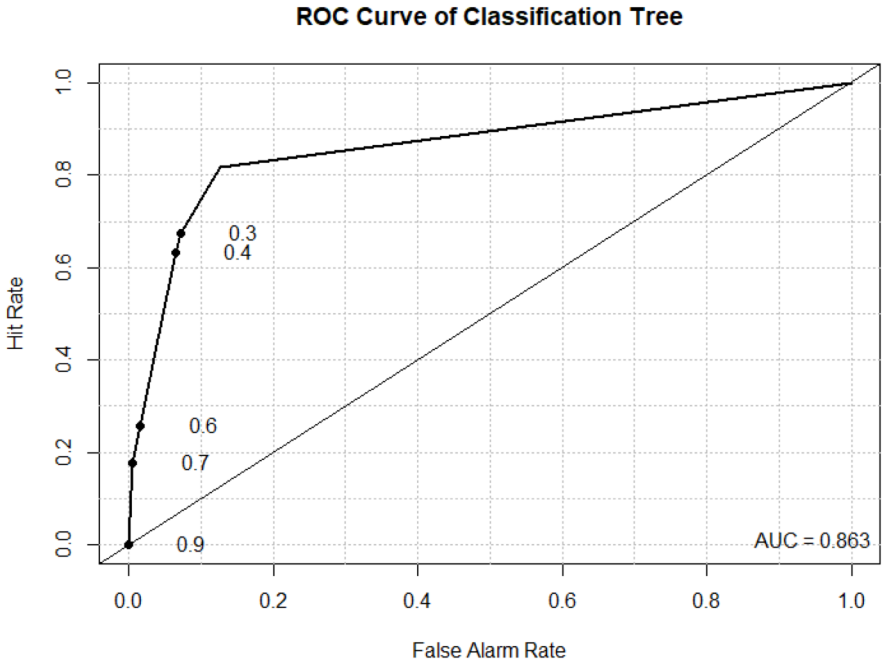


Lasso

- Variables shrunk to 0:
  - Administrative, Administrative_Duration
  - Informational, Information_Duration
  - ProductRelated_Duration
  - BounceRate
  - SpecialDay (!)
  - Weekend
- Test Error: 11.8%
- Type I Error Rate: 5.2%
- Type II Error Rate: 50.3%
- Power: 49.7%

Classification Tree

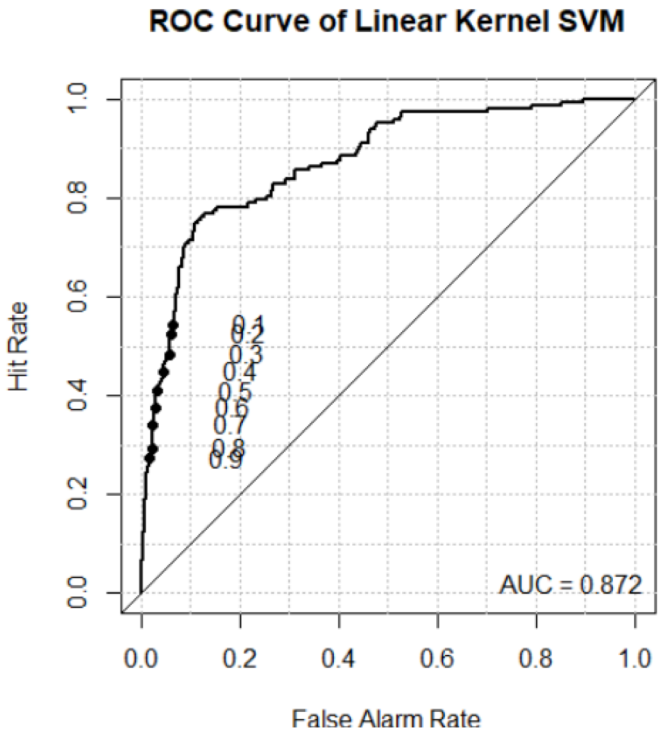
- Test Error: 10.9%
- Type I Error: 6.4%
- Type II Error: 36.7%
- Power: 63.3%
- AUC: 0.863



SVM - Linear Kernel

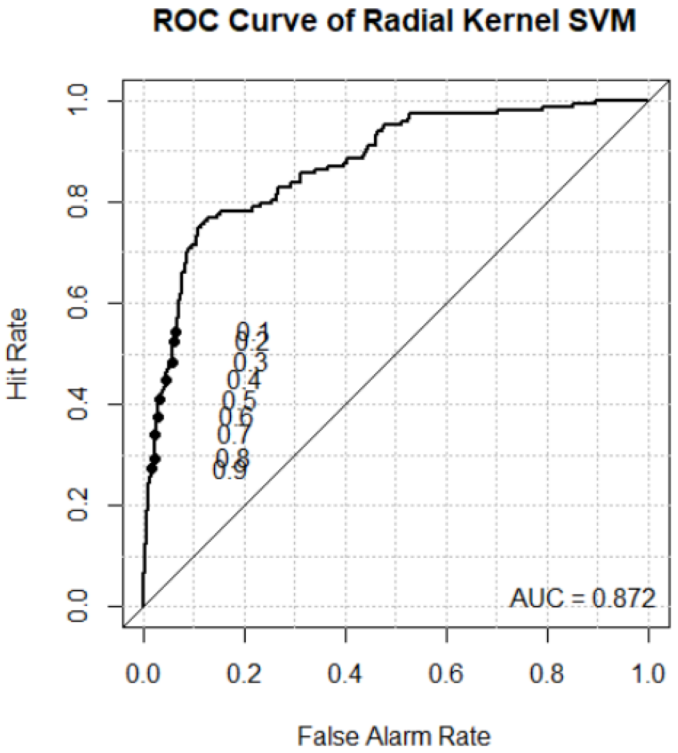
Tuning the model with multiple cost values

- cost = 0.01, 0.1, 0.5, 1, 5, best at 1
- 915 support vectors(4000 observations in training set)
- Type I: 7.03%
- AUC: 0.8722396



SVM - Radial Kernel

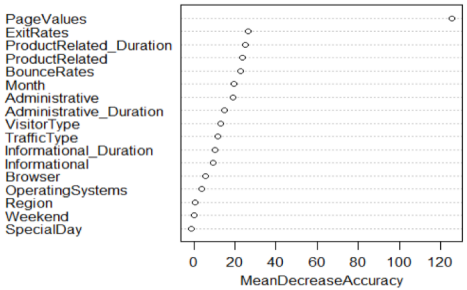
- Running extremely slow
- Tuning the model with multiple cost and gamma values
  - cost =0.01, 0.1, 1, 5, best at 1
  - gamma = 0.001, 0.01, 0.1, 0.5, 1, best at 0.1
  - 1046 support vectors
  - overall error rate: 11%
  - Type I: 5.6%
  - AUC: 0.8722396

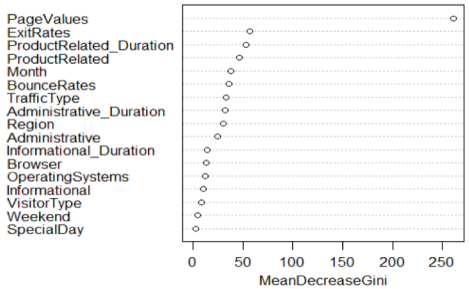


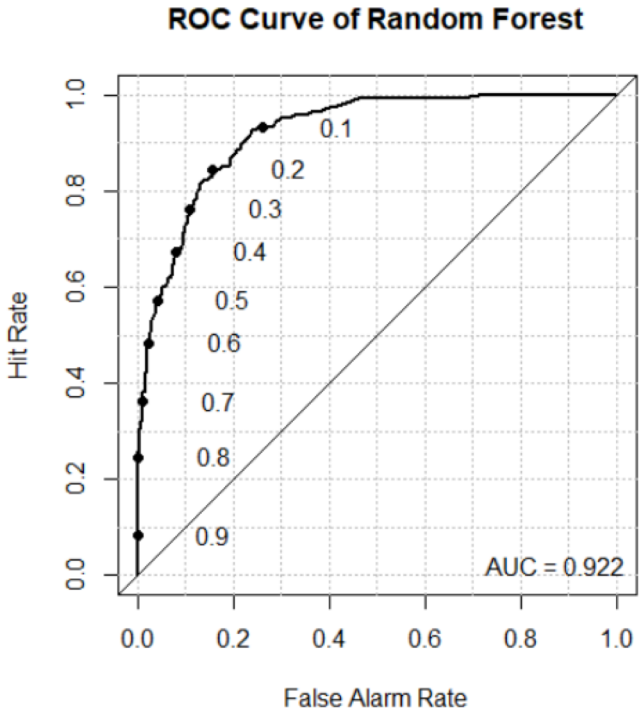
Random Forest

- Start with 500 trees
- Tuning the model with OOB error - find the number of trees that produces the minimum OOB error rate
    - 296 trees
- Rebuild the tree with the optimal number of trees
- Predict with the test set
- Type I Error: 4.22%
- AUC: 0.922327

Random Forest



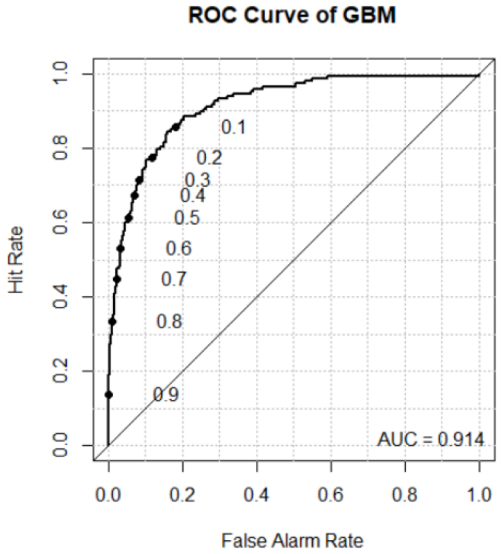




Gradient Boosting Machines

- Tuning the hyper parameters - 81 combinations
- Rebuild the model based on the best parameters
- Type I error: 5.74%
- AUC: 0.9136062

```
hyper_grid <- expand.grid(
  shrinkage = c(.01, .05, .1),
  interaction.depth = c(3, 5, 7),
  n.minobsinnode = c(5, 7, 10),
  bag.fraction = c(.65, .8, 1),
  optimal_trees = 0,
  min_acc = 0
)
```



Results

Model	Overall Error	Type I Error	Type II Error	AUC
Lasso	0.118	0.05158265	0.5034014	0.9015799
Logistic Regression	0.128	0.05627198	0.5442177	0.8856776
Classification Decision Tree	0.109	0.06447831	0.3673469	0.8630484
Linear SVM	0.121	0.07033998	0.414966	0.8722396
Radial SVM	0.11	0.05627198	0.4217687	0.8722396
Random Forest	0.099	0.04220399	0.4285714	0.922327
GBM	0.106	0.05744431	0.3877551	0.9136062

Summary

- Type I Error
  - We predict customers make a purchase, but they don't
- Type II Error
  - We predict customers don't make a purchase, but they do
- AUC
  - Overall prediction accuracy (combination of type I + type II)

Most Important Factors

1. Page Values
   The average value for a web page that a user visited before completing an e-commerce transaction

2. Exit Rates
   The percentage of all pageviews to the page

3. Product Related Duration
   The total time the visitor spent on product related pages

4. Product Related
   The number of product related pages visited during the session

5. Bounce Rate
   The % of visitors who enter the site from that page and then leave without triggering any other requests to the analytics server during that session

Recommendation

- Page Value is the most important factor, keep customers clicking through the website
  - Create additional content for users to browse
  - Make sure website is easy to navigate
  - Create algorithm for blog posts/social content based on user navigation history
  - Helpful input for A/B testing and structure segments
- Reduce Exit Rates by incentivizing users to stay on the site (similar to Page Value)
- Continue to improve Product Recommendation algorithm

Bonus slides - Correlation Matrix