Predicting Brand Preferences With Demographic Data and Machine Learning Algorithms

Benjamin P. Fauber, Ph.D.
Austin, Texas USA

September 15, 2016

Goal: Predict Customer Preference for Brand of Laptop Based on Other Demographic Data

1) Provided a complete set of customer demographic data and laptop brand preference.

2) Build, cross-validate, and refine classification models on the complete dataset.

3) Test the refined classification models on hold-out data.

4) Apply the best model to new demographic data and predict customer preference for a certain brand of laptop (Acer or Sony).

[10,000 x 7] data matrix, ARFF format

Retailer Provided Data

age (20-80), annual salary ($20-150k), education level (HS-PhD), primary car (20 types), US region (8 options), available credit ($0-500k), and laptop brand preference (Acer or Sony)
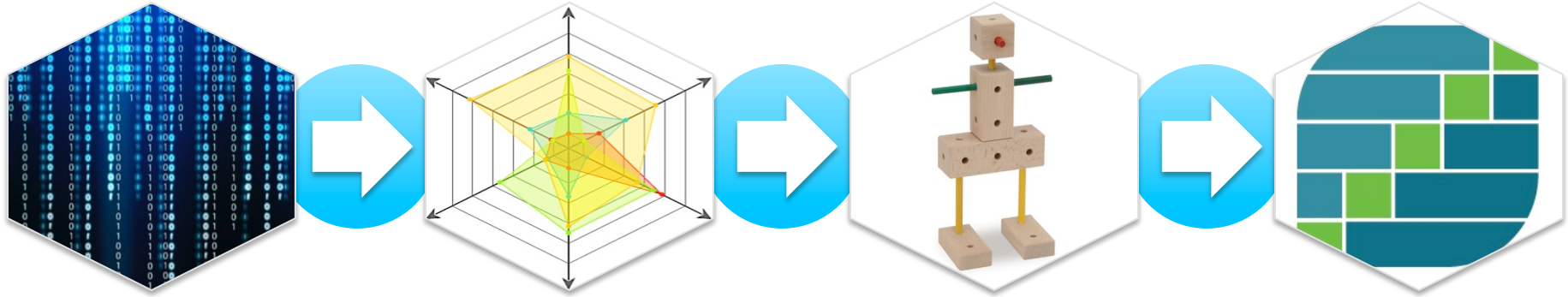
Approach

70,000 data points

5 analyses

Conducted Analysis with R and Employed Machine Learning Algorithms

1 **Import**

R Language

Clean Data

2 **Analyze**

R Language

Plot Data

3 **Model**

R Language

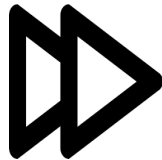*Machine Learning*

4 **Predict**

R Language

Report

Cleaned Data: Transformed Columns to Numeric and Factors and Removed NULL Values
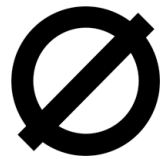
Retailer provided data

Fairly clean, but required some additional cleaning to facilitate analysis

R *as.numeric()* or *as.factor()*

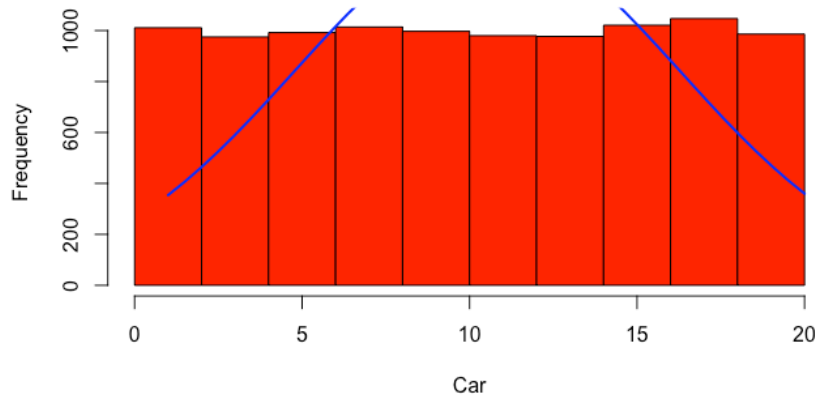**Values to Numeric or Factor**

Applied to: All
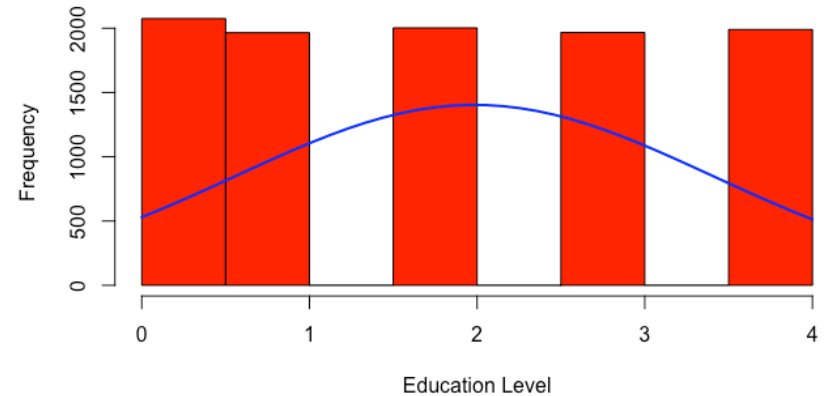
R *na.exclude()*

Removed NULL Values

Applied to: All

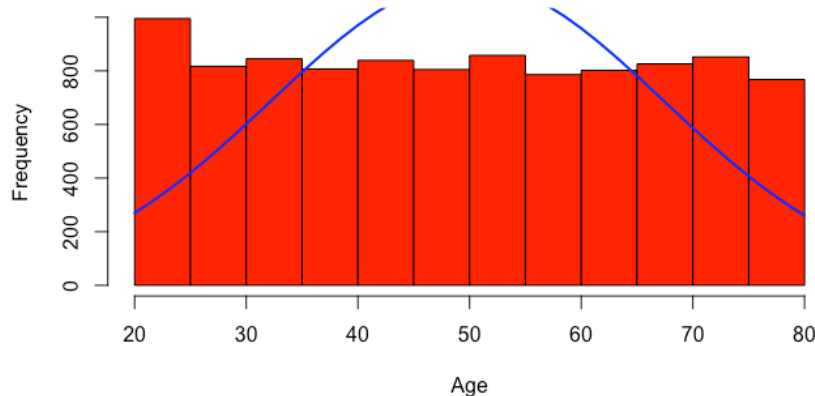# Plotting the Cleaned Data Revealed a Reasonable Distribution of Features – No Dominant Values

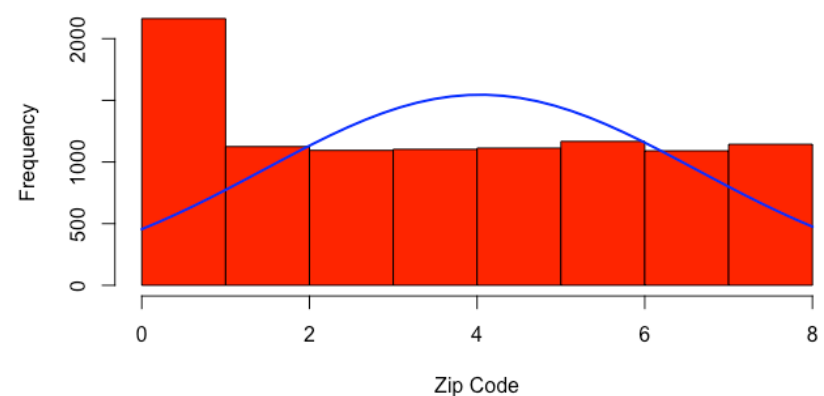# Split the Data Into Two Sets: Training and Hold-Out – Optimized on Training, Evaluated Models Against Both
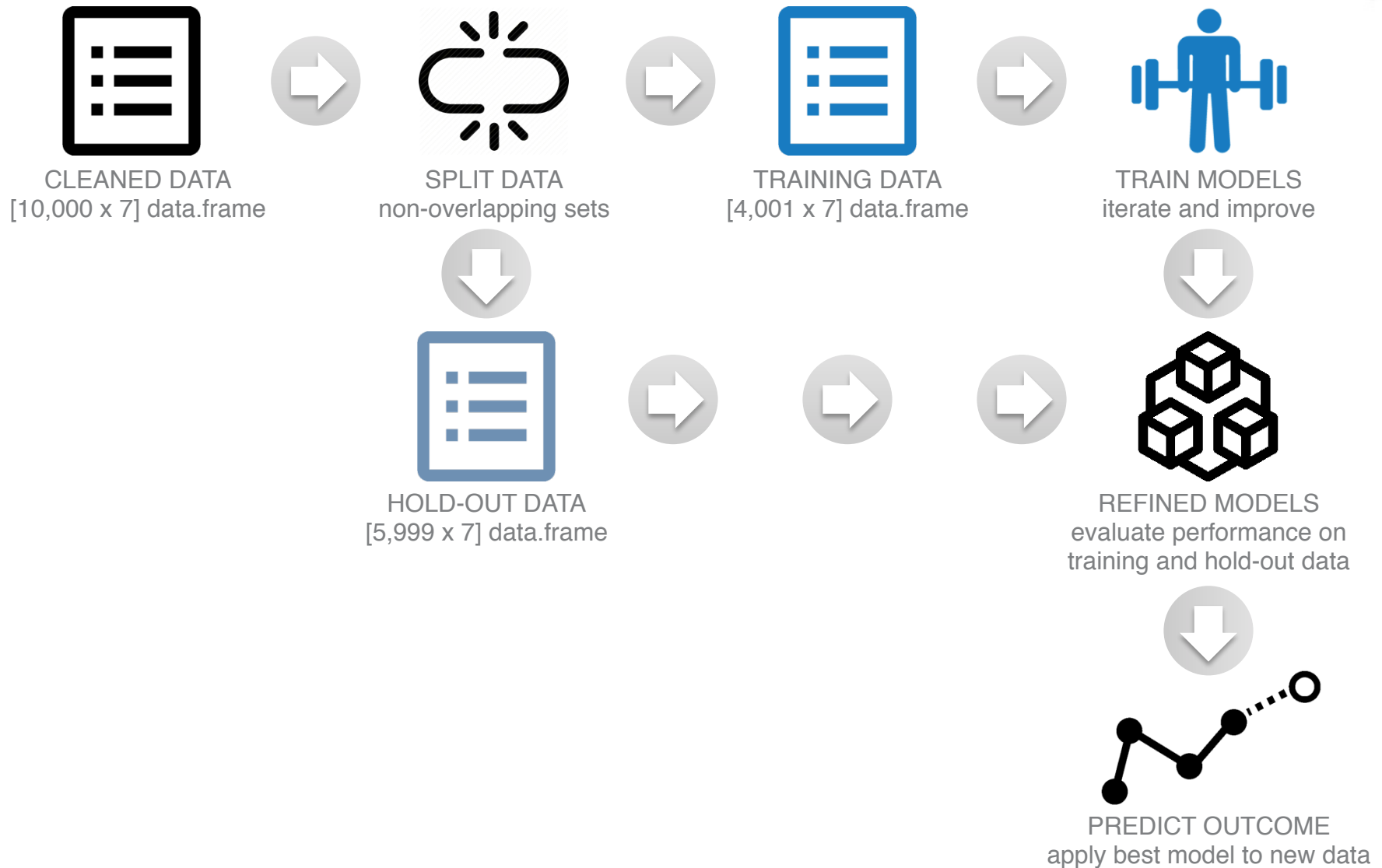
CLEANED DATA
[10,000 x 7] data.frame

SPLIT DATA
non-overlapping sets

TRAINING DATA
[4,001 x 7] data.frame

TRAIN MODELS
iterate and improve

HOLD-OUT DATA
[5,999 x 7] data.frame

REFINED MODELS
evaluate performance on
training and hold-out data

PREDICT OUTCOME
apply best model to new data

RandomForest Classifier

```
R RFfit <- train(brand ~., data = training,
method = "rf", metric = "Accuracy", trControl =
fitControl, ntree = 500, tuneGrid = rfGrid6,
importance = TRUE)
```


e1071 Support Vector Classifier

```
R e1071SVMopt <- tune.svm(brand ~ ., data =
training, metric = "Accuracy", gamma =
svm5Grid, cost = svm5Grid, kernel = "radial",
trControl = fitControl)
```
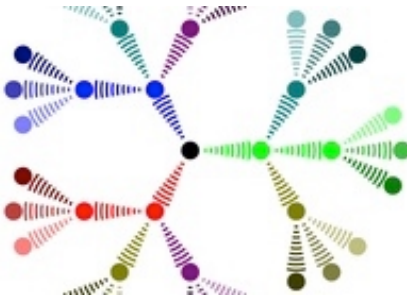
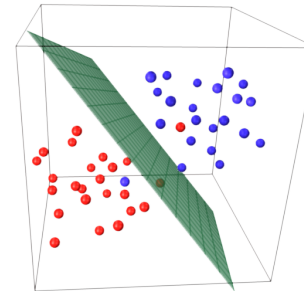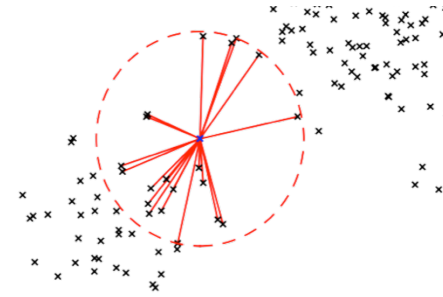

c5.0 Statistical Classifier

```
R c50fit <- train(brand ~., data = training,
metric = "Accuracy", trControl = fitControl,
method = "C5.0", tuneLength = 10, verbose =
FALSE)
```


kNN Nearest-Neighbor Classifier

```
R kNNfit <- train(brand ~., data = training,
method = "knn", metric = "Accuracy", trControl
= fitControl, tuneLength = 20, preProc =
c("range"))
```

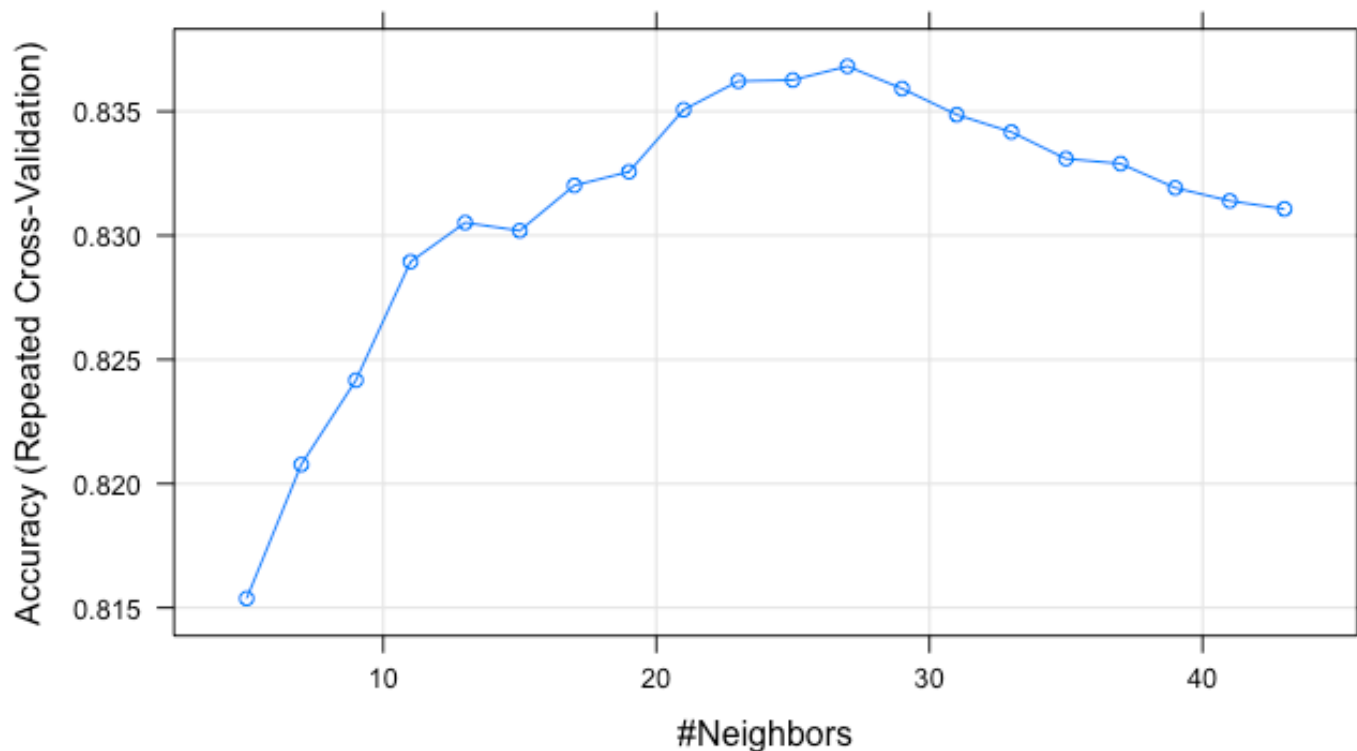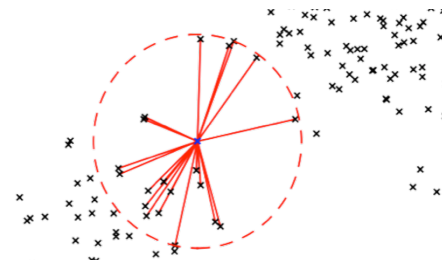## kNN Nearest-Neighbor Classifier

```
R kNNfit <- train(brand ~., data = training, method =
"knn", metric = "Accuracy", trControl = fitControl,
tuneLength = 20, preProc = c("range"))
```

Calculated Classification Model Metrics of Predicted Values vs. Actual Training Set Values For All Approaches

Model Name	Accuracy	Precision	Recall	Sensitivity	Specificity
RandomForest	1	1	1	1	1
c5.0	0.93	0.92	0.90	0.90	0.95
e1071 SVM radial	0.91	0.85	0.92	0.92	0.90
kNN	0.86	0.83	0.81	0.81	0.90
e1071 SVM poly	0.69	0.78	0.24	0.24	0.96

Values derived from Confusion Matrix for each model, comparing predicted vs. training values

Accuracy = correctly categorized
Precision = fraction of predicted that are actually in that class
Recall = fraction of actual that detected by the model
Sensitivity = true-positive rate
Specificity = true-negative rate

Even With 10-fold Cross-Validation to Limit Overfitting it Appears RandomForest and c5.0 May Still Overfit

Calculated Classification Model Metrics of Predicted Values vs. Actual Training Set Values For All Approaches

Model Name	Accuracy	Precision	Recall	Sensitivity	Specificity
RandomForest	1	1	1	1	1
c5.0	0.93	0.92	0.90	0.90	0.95
e1071 SVM radial	0.91	0.85	0.92	0.92	0.90
kNN	0.86	0.83	0.81	0.81	0.90
e1071 SVM poly	0.69	0.78	0.24	0.24	0.96

Further validating models against hold-out data will examine if certain models overfit the training data

RandomForest and c5.0 Classification Models Performed Very Well on Hold-Out Data – Not Overfit

Calculated Classification Model Metrics of Predicted Values vs. Training and Hold-Out Values For All Approaches
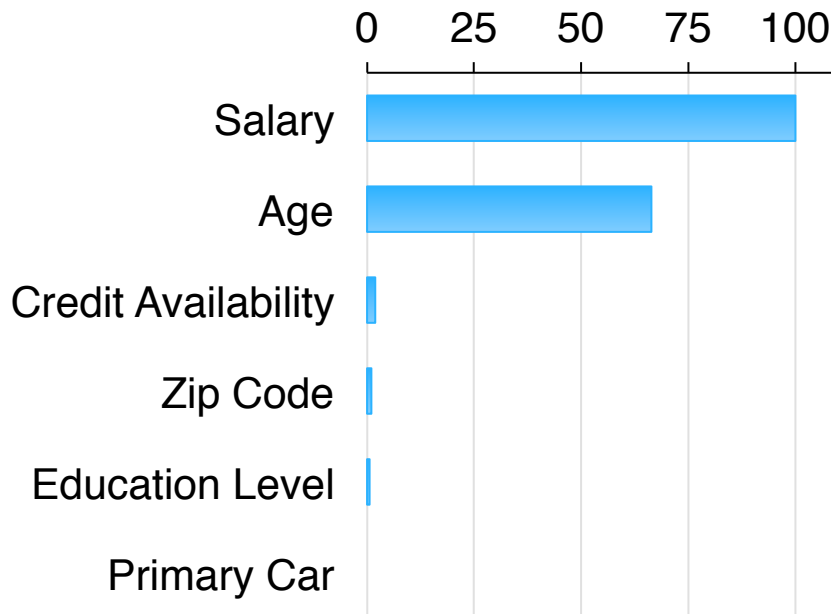
Model Name (dataset)	Accuracy	Precision	Recall	Sensitivity	Specificity
RandomForest (train)	**1**	1	1	1	1
RandomForest (hold-out)	**0.92**	0.90	0.89	0.89	0.94
c5.0 (train)	**0.93**	0.92	0.90	0.90	0.95
c5.0 (hold-out)	**0.92**	0.91	0.88	0.88	0.95
e1071 SVM radial (train)	0.91	0.85	0.92	0.92	0.90
e1071 SVM radial (hold-out)	0.86	0.80	0.86	0.86	0.87
kNN (train)	0.86	0.83	0.81	0.81	0.90
kNN (hold-out)	0.84	0.80	0.77	0.77	0.88
e1071 SVM poly (train)	0.69	0.78	0.24	0.24	0.96
e1071 SVM poly (hold-out)	0.67	0.72	0.19	0.19	0.95

Salary and Age of Customers Were the Most Influential Fields in the RandomForest and c5.0 Models
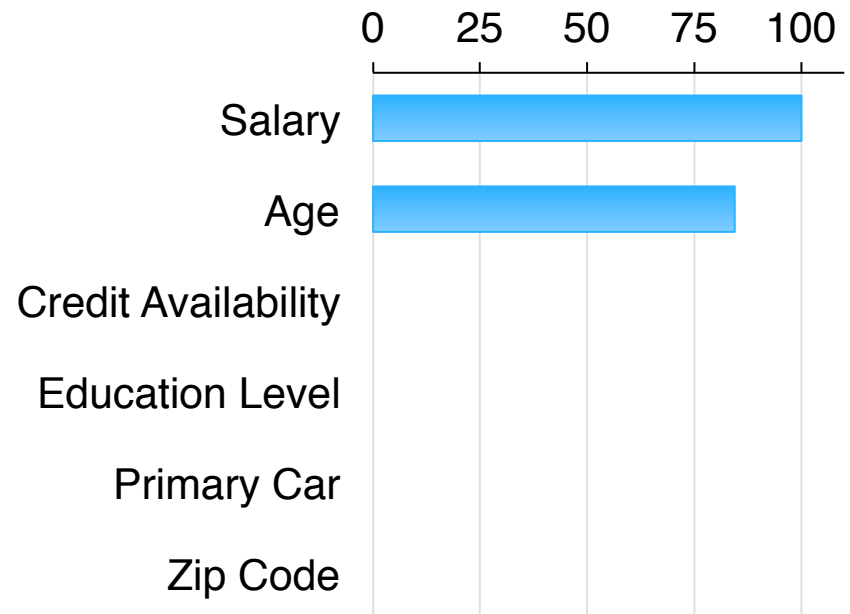
Influential RandomForest Fields

Salary and Age were the most influential in the model – all values are normalized to 100.

	0	25	50	75	100
Salary					
Age					
Credit Availability					
Zip Code					
Education Level					
Primary Car					

Influential c5.0 Fields

Only Salary and Age were used by the model – all values are normalized to 100.

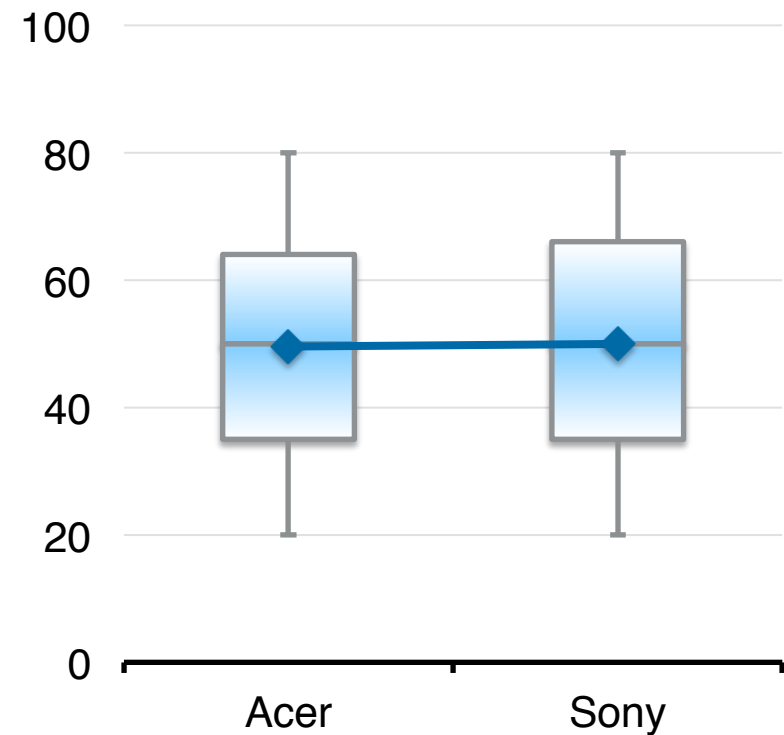|	0	25	50	75	100
Salary					
Age					
Credit Availability					
Education Level					
Primary Car					
Zip Code					

Salary vs. Brand Preference

Sony customers had higher mean, median, and max salaries than Acer customers in the Cleaned Data.
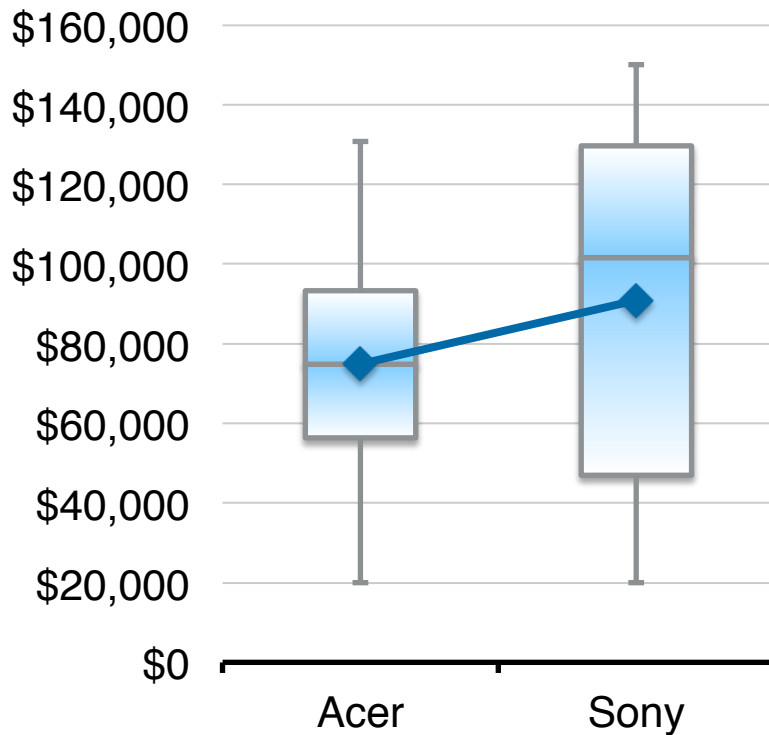
Age vs. Brand Preference

All age statistics were nearly identical for the Acer and Sony customer groups in the Cleaned Data.
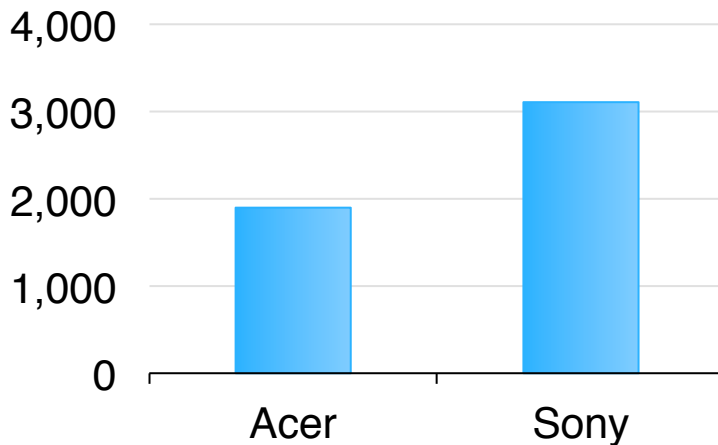
Applied Validated RandomForest and c5.0 Models to New Data to Predict Customer Brand Preferences
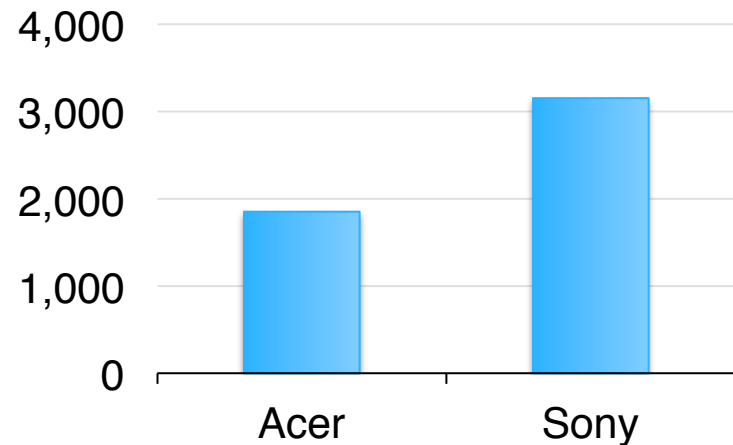
RF Model Brand Prediction

Application of model to new data shows preference for Sony over Acer.

c5.0 Model Brand Prediction

Application of model to new data shows preference for Sony over Acer.


```
NewData # new data, missing brand preferences, 5000 x 7 dataframe

# RF model, optimized with mtry = 3 and ntree = 500
RFfitNew <- data.frame(predict(class6fit, NewData, interval="predict",
level=0.95, na.action=na.omit))

# c5.0 model, optimized with trials = 10, model = tree, winnow = T
c50fitNew <- data.frame(predict(class2fit, NewData, interval = "predict", level
=0.95, na.action=na.omit))
```

Recommendations for Future Data Collection Campaigns to Inform Customer Brand Preferences

1) Future data collection campaigns should gather the customer age and salary as the first questions (most important fields in current analyses).

2) Continue to collect additional customer information such as Credit Availability, Zip Code (US Region), and Education Level as it may be important in other brand preference analyses.

3) Important to collect some information on customer brand preferences in future surveys to validate future demographic data and survey models (allows for creation of future training and hold-out datasets).

4) Primary car was not an important feature in the top two models – might consider making this a later question or dropping this question/field.

Benjamin P. Fauber, Ph.D.
Austin, Texas USA

https://github.com/BFauber