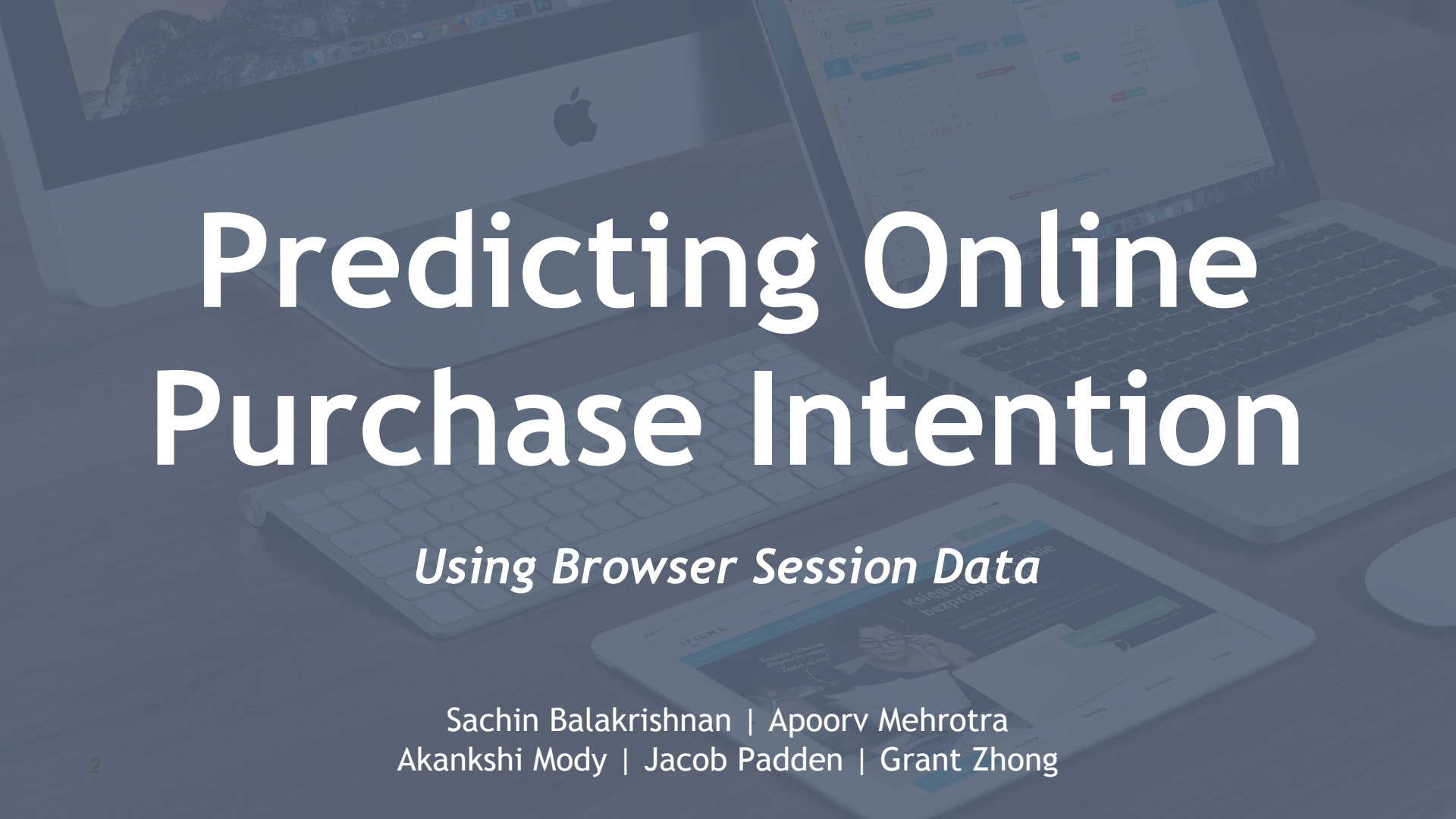


Predicting Online Purchase Intention

Using Browser Session Data

Sachin Balakrishnan | Apoorv Mehrotra
Akankshi Mody | Jacob Padden | Grant Zhong

DATASET DESCRIPTION

Dataset Description

12,330 sessions

1-year period

10 numerical attributes

8 categorical attributes

Level of Data - Browser Session

1. A set of **hits** triggered by a user
2. A hit is a user interaction (pageview, screenview, event, transaction) that sends data to GA server (GIF Request)
3. A user can generate >1 sessions in a day

Dataset Attributes

1. **Revenue**
2. Administrative
3. Administrative Duration
4. Informational
5. Informational Duration
6. Product Related
7. Product Related Duration
8. Exit Rate
9. Bounce Rate
10. Page Value
11. Special Day (duration between the order date and delivery date)
12. Weekend
13. Month of the year
14. Operating system
15. Browser
16. Region
17. Traffic type
18. Visitor type (returning or new visitor)

CLASSIFICATION MODELS

Logistic Regression

- Binary Classification Problem.
- Dependent Variable - Whether the user will generate revenue for the website?
- Seventeen predictor variables – 10 numeric and 7 categorical.
- 12330 observations

Overall Accuracy	88.49%
Sensitivity (True Positive Rate)	*38.39%*
Specificity (True Negative Rate)	97.61%
Precision	75.03%
F1 - Score	0.513

Model Performance – NOT GOOD !

- Skewed Dependent variable.
- Ratio of Positive to Negative Observations – 85% to 15%
- Always predict FALSE – you got 85% accuracy !!!
- Too many FALSE NEGATIVES
- LOW AUC - 0.683

CONFUSION MATRIX		OBSERVED	
		FALSE	TRUE
PREDICTED	FALSE	7115	819
	TRUE	174	523

Resampling Methods

Down-Sampling

Under-sample the majority class instances from the training data

Up-Sampling

Duplicate the minority instances in the training data

SMOTE

Synthetic Minority Oversampling Technique
Nearest neighbour approach to draw artificial samples

ROSE

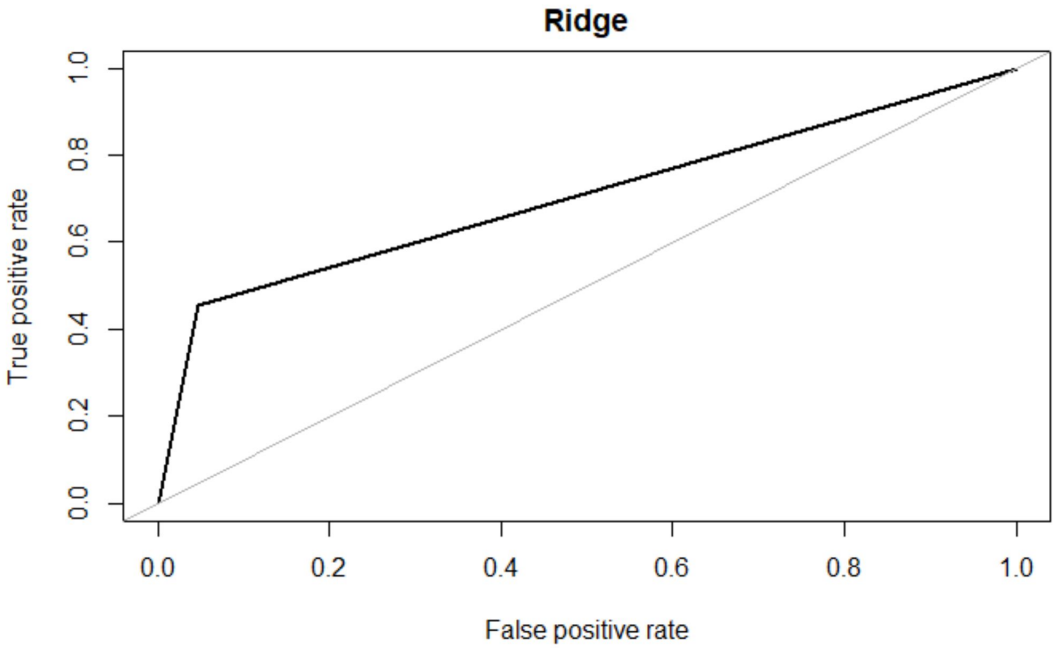
Random Over-Sampling Examples
Smoothed bootstrapping to draw artificial samples from feature space

Resampling with 10 fold CV

SMOTE yielded the best results in-terms of accuracy and AUC value.

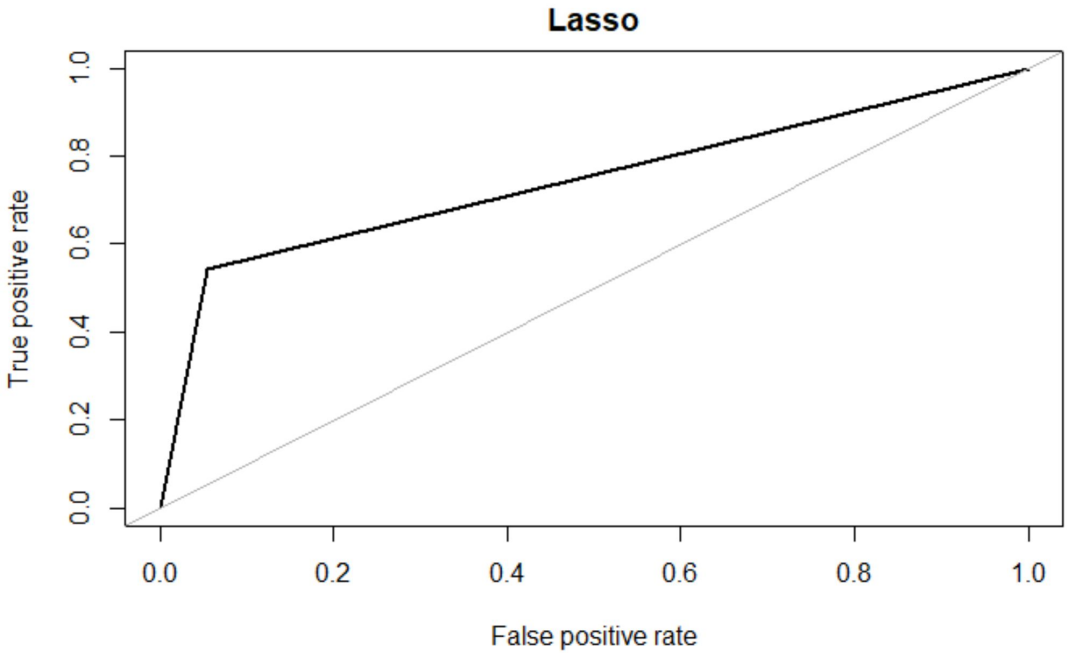
	Total Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC
Down-Sampled	0.7949	0.7608	0.829	0.8161	0.7875	0.795
Up-Sampled	0.8179	0.7797	0.8561	0.8442	0.8107	0.818
SMOTE	*0.853*	*0.7635*	*0.9201*	*0.8775*	*0.8165*	*0.842*
ROSE	0.8168	0.7616	0.8676	0.8412	0.7994	0.815

Ridge Regression



Using SMOTE Resampling	
Min Lambda	0.0247
Accuracy	0.872
Sensitivity	0.456
Specificity	0.953
Precision	0.656
F1 Score	0.538
AUC	0.705

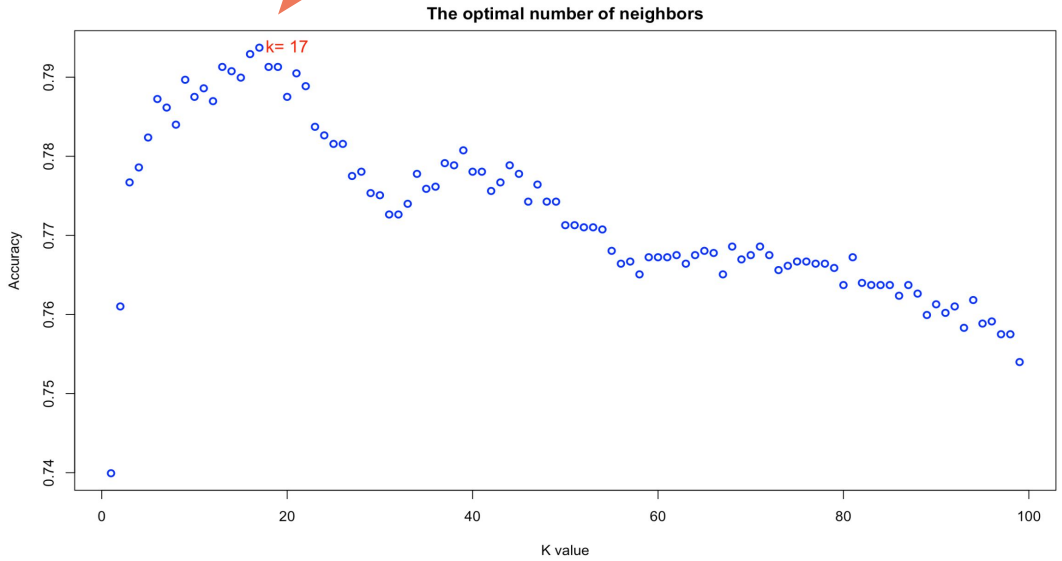
Lasso Regression



Using SMOTE Resampling	
Min Lambda	0.00058
Accuracy	0.879
Sensitivity	0.544
Specificity	0.945
Precision	0.659
F1 Score	0.596
AUC	0.744

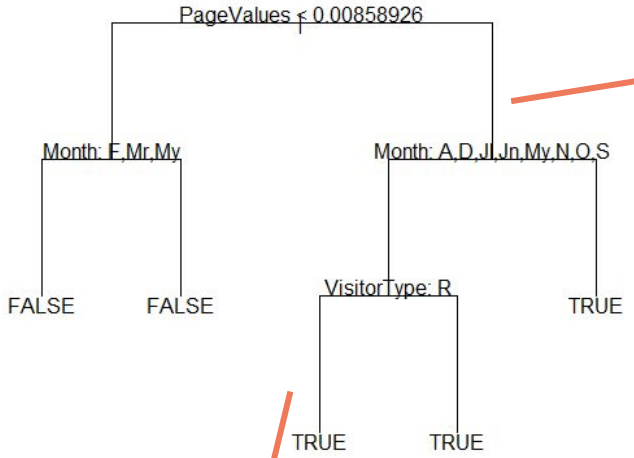
K-Nearest Neighbor w/ SMOTE

k=17 gave highest accuracy

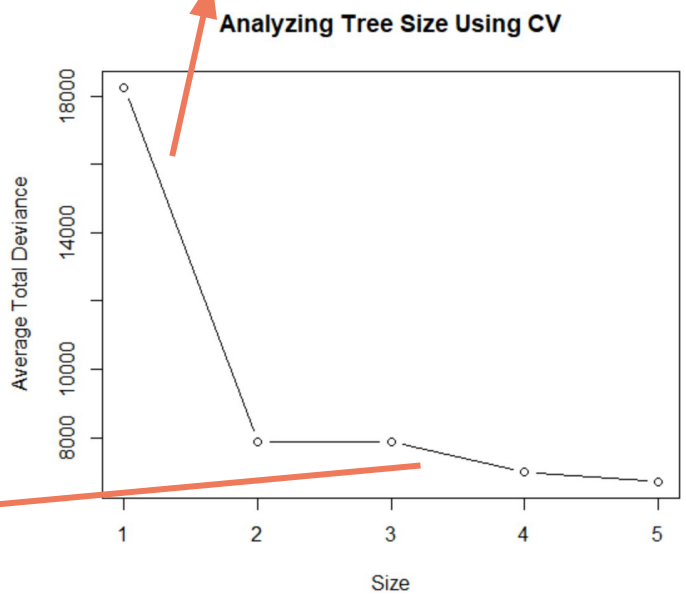


10 Fold CV with SMOTE Resampling	
Accuracy	0.793
Sensitivity	0.389
Specificity	0.928
Precision	0.389
F1 Score	0.478

Decision Tree



PageValues extremely important!

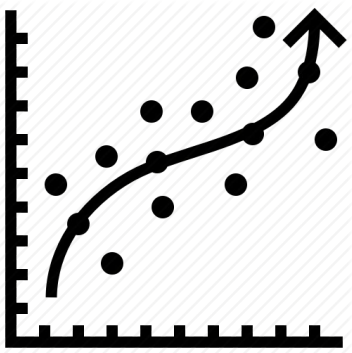


Other variables fine tune probabilities

10 Fold CV with SMOTE Resampling	
Accuracy	0.876
Sensitivity	0.785
Specificity	0.893
Precision	0.575
F1 Score	0.662
AUC	0.839

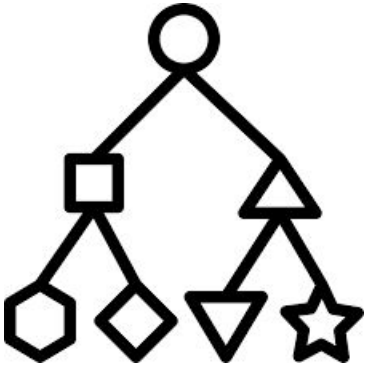
Random forest - 2 approaches

Regression



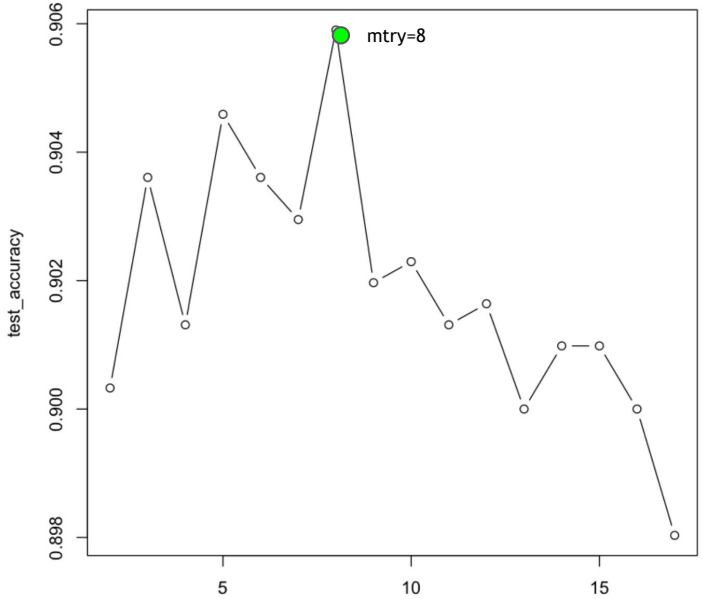
1. Tune for mtry
2. Tune for cutoff
3. Compare metrics
4. Check variable importance

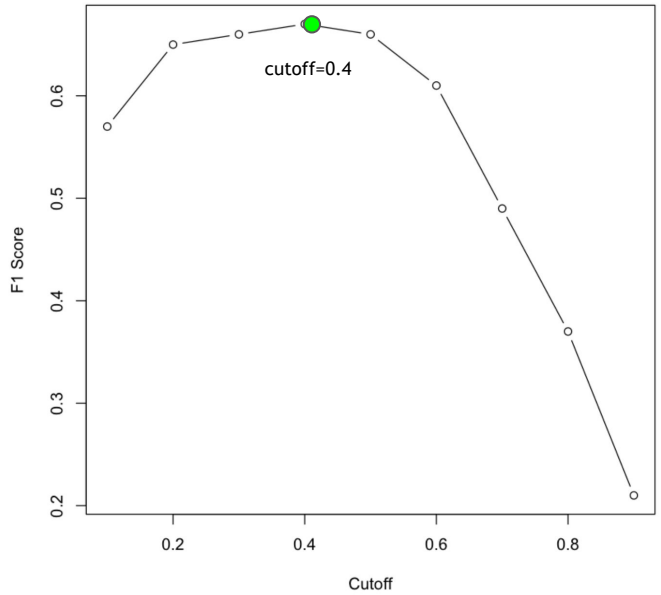
Classification



1. Tune for mtry
2. Compare metrics
3. Check ROC
4. Check variable importance

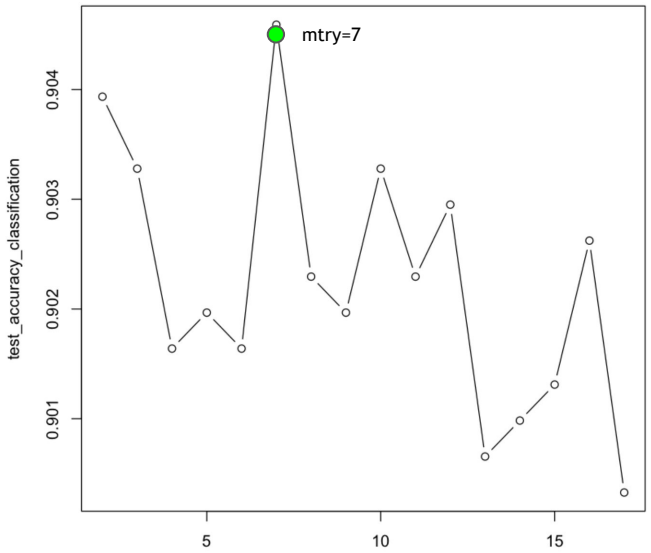
Random forest - Regression

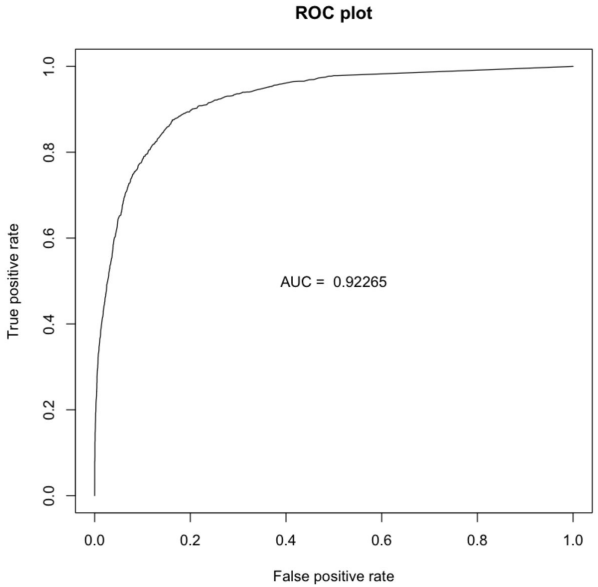




Accuracy	Sensitivity (Recall)	Specificity	Precision	F1-Score
0.89	0.68	0.94	0.66	0.67

Random forest - Classifier

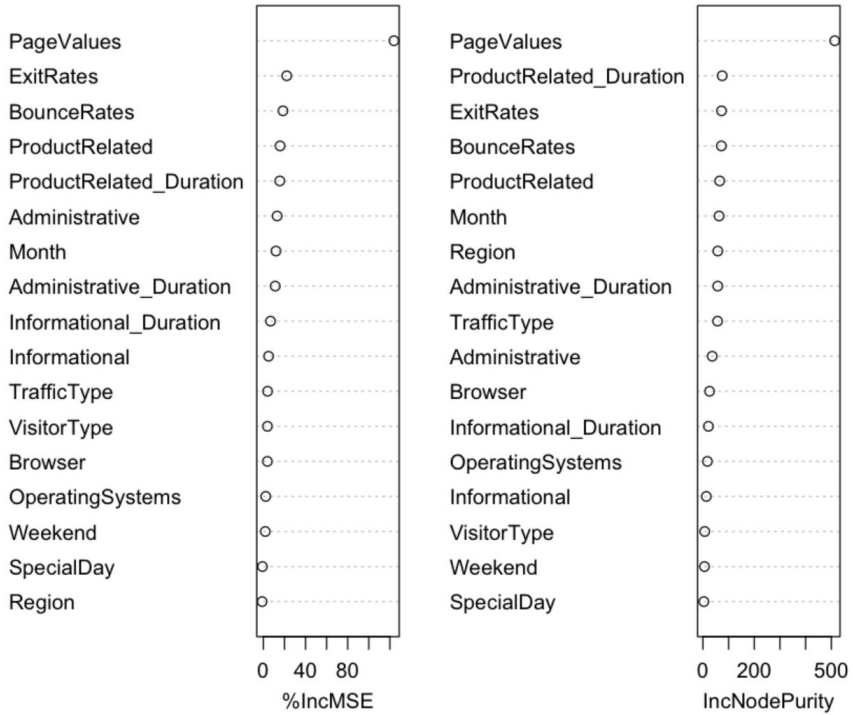




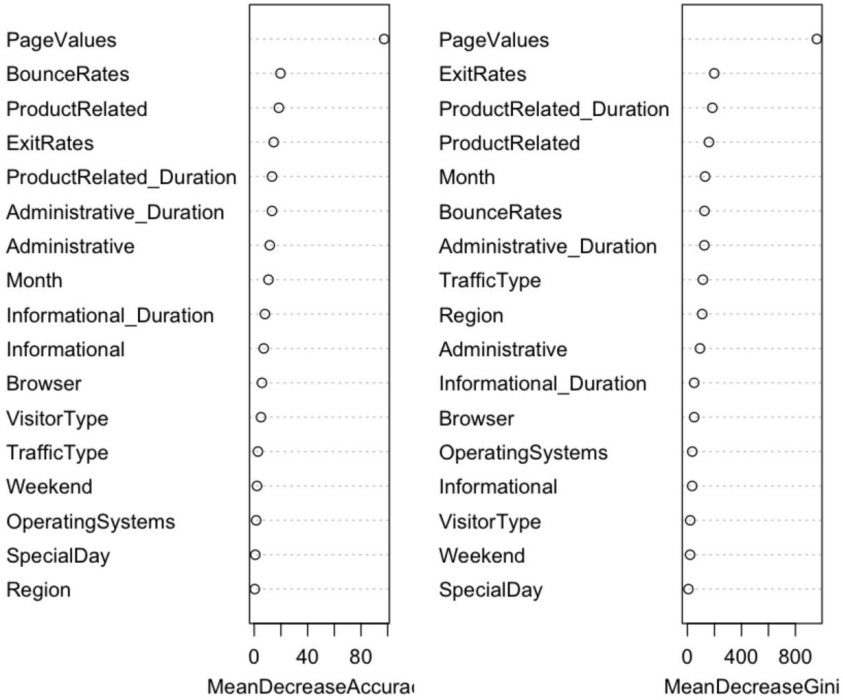
Accuracy	Sensitivity (Recall)	Specificity	Precision	F1-Score	AUC
0.904	0.63	0.95	0.71	0.66	0.922

Random forest - Important Variables

Regression



Classification



CONCLUSION

Model Selection

Model	*Accuracy*	*F1 Score*
Logistic Regression	0.88	0.82
Random Forest	0.89	0.67
Decision Tree	0.88	0.66
LASSO Regression	0.88	0.60
Ridge Regression	0.87	0.54
KNN	0.79	0.48

Thanks for Watching

We hope you enjoyed watching the presentation
as much as we enjoyed making it.


QUESTIONS?

Acknowledgments

- C. Okan Sakar Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Bahcesehir University, 34349 Besiktas, Istanbul, Turkey
- Yomi Kastro Inveon Information Technologies Consultancy and Trade, 34335 Istanbul, Turkey

Source: https://www.kaggle.com

Appendix

	Estimate	Std. Error	z value	Pr(>|z|)	
Informational	7.828e-02	2.024e-02	3.868	0.000110	***
ExitRates	-1.495e+01	1.749e+00	-8.547	< 2e-16	***
PageValues	1.228e-01	2.972e-03	41.322	< 2e-16	***
MonthDec	-6.862e-01	1.410e-01	-4.867	1.13e-06	***
MonthFeb	-1.632e+00	3.831e-01	-4.260	2.05e-05	***
MonthMar	-6.668e-01	1.394e-01	-4.784	1.72e-06	***
MonthMay	-6.837e-01	1.340e-01	-5.104	3.33e-07	***
MonthNov	6.824e-01	1.288e-01	5.298	1.17e-07	***
Browser6	-1.335e+00	2.906e-01	-4.594	4.36e-06	***
Region4	-4.047e-01	8.814e-02	-4.591	4.41e-06	***
Region5	-5.647e-01	1.586e-01	-3.562	0.000369	***
TrafficType2	2.729e-01	6.889e-02	3.961	7.47e-05	***
TrafficType5	5.811e-01	1.566e-01	3.710	0.000207	***
TrafficType8	5.700e-01	1.357e-01	4.199	2.68e-05	***
TrafficType10	5.369e-01	1.200e-01	4.472	7.73e-06	***
TrafficType11	6.193e-01	1.606e-01	3.857	0.000115	***
TrafficType13	-6.856e-01	1.393e-01	-4.922	8.56e-07	***
TrafficType20	1.002e+00	2.151e-01	4.658	3.20e-06	***