PREDICTION OF AIRBNB BOOKING DESTINATION

AMEYA S THOMBRE, VARUN SRIVASTAVA, ISHAN DHAWAN

TEAM DELTA FORCE

PURDUE UNIVERSITY

Introduction & Background

- This project involves prediction of new user booking destination using Machine Learning Algorithms on the Airbnb dataset.
- This will enable Airbnb to pay specific attention to the most important factors in order to bolster their bookings.
- Accurate predictions can also prove to be useful in helping Airbnb optimize their strategies to expand their user base and maximize their profits.

PURDUE UNIVERSITY

DATA SECTION

- Data available from 2012-2014 for users in USA.
- Available Parameters

Input Variable	Description
Date Account Created	Date on which the User created the account
Date First Booking	Date on which the User booked for the first time
Gender	Gender of the user
Age	Age of the User
Language	The language of the Website which the Customer used
Affiliate Channel	Kind of Paid marketing strategy which led the customer to sign up
Affiliate Provider	The kind of Marketing firm which served as the Affiliate channel
Signup App	Type of App used to Sign Up
First Device Type	Type of Device used to Sign up for the first time
First Browser	Type of Browser used to Sign up for the first time
Secs Elapsed	Amount of time spent on the website, taken from the sessions dataset.
Destination Distance	The distance of the destination country from the US. Taken from the Countries dataset.

PURDUE UNIVERSITY

Output Variable

- Categorical Output Variable (Destination Country)

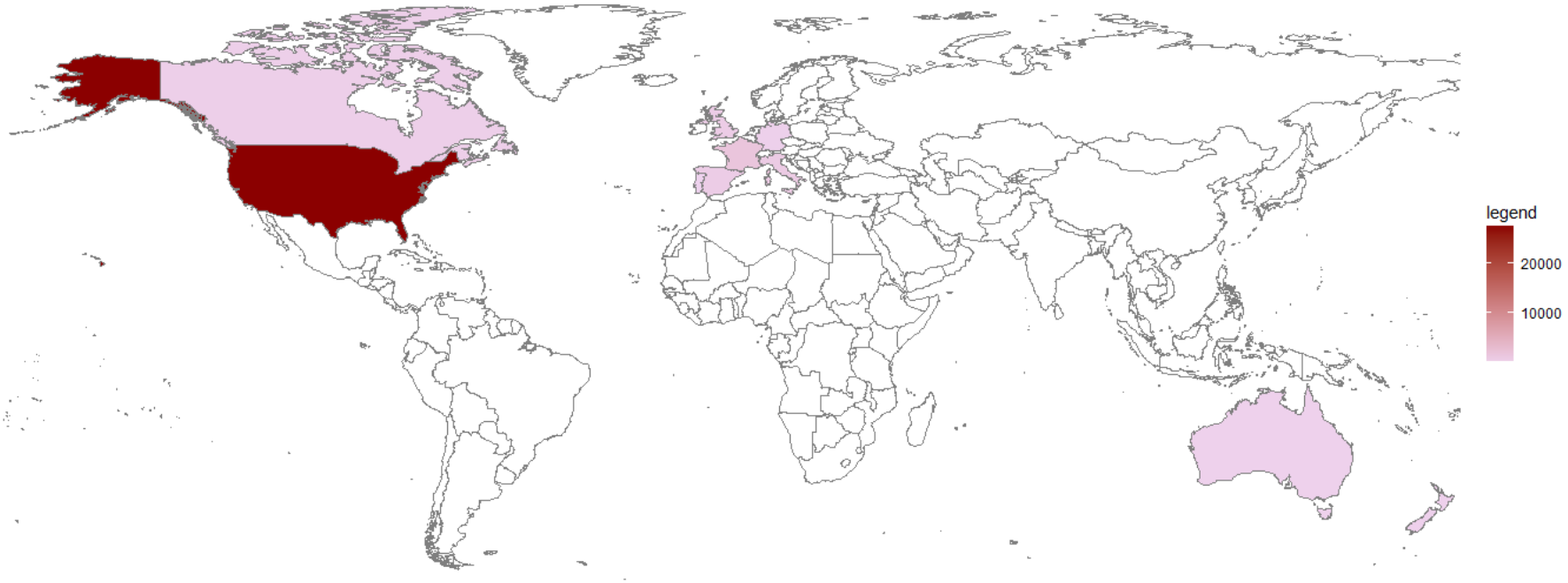
Sr. No	Destination Country
1.	Australia (AU)
2.	Canada (CA)
3.	Germany(DE)
4.	Spain (ES)
5.	France (FR)
6.	Great Britain (GB)
7.	Italy (IT)
8.	NDF (Did not book)
9.	Netherlands (NL)
10.	other
11.	Portugal (PT)
12.	United States (US)

PURDUE UNIVERSITY

EXPLORATORY DATA ANALYSIS

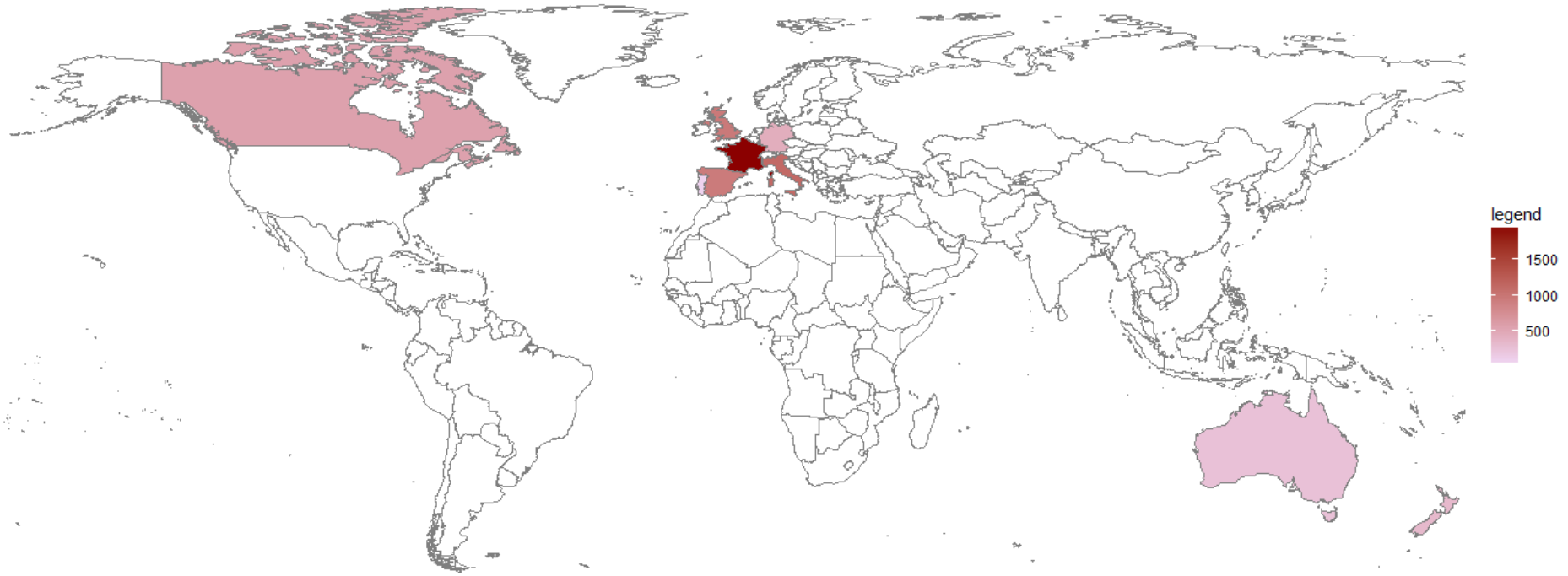
Number of bookings (Country-wise)

- Mapping number of bookings by destination country on the complete dataset.



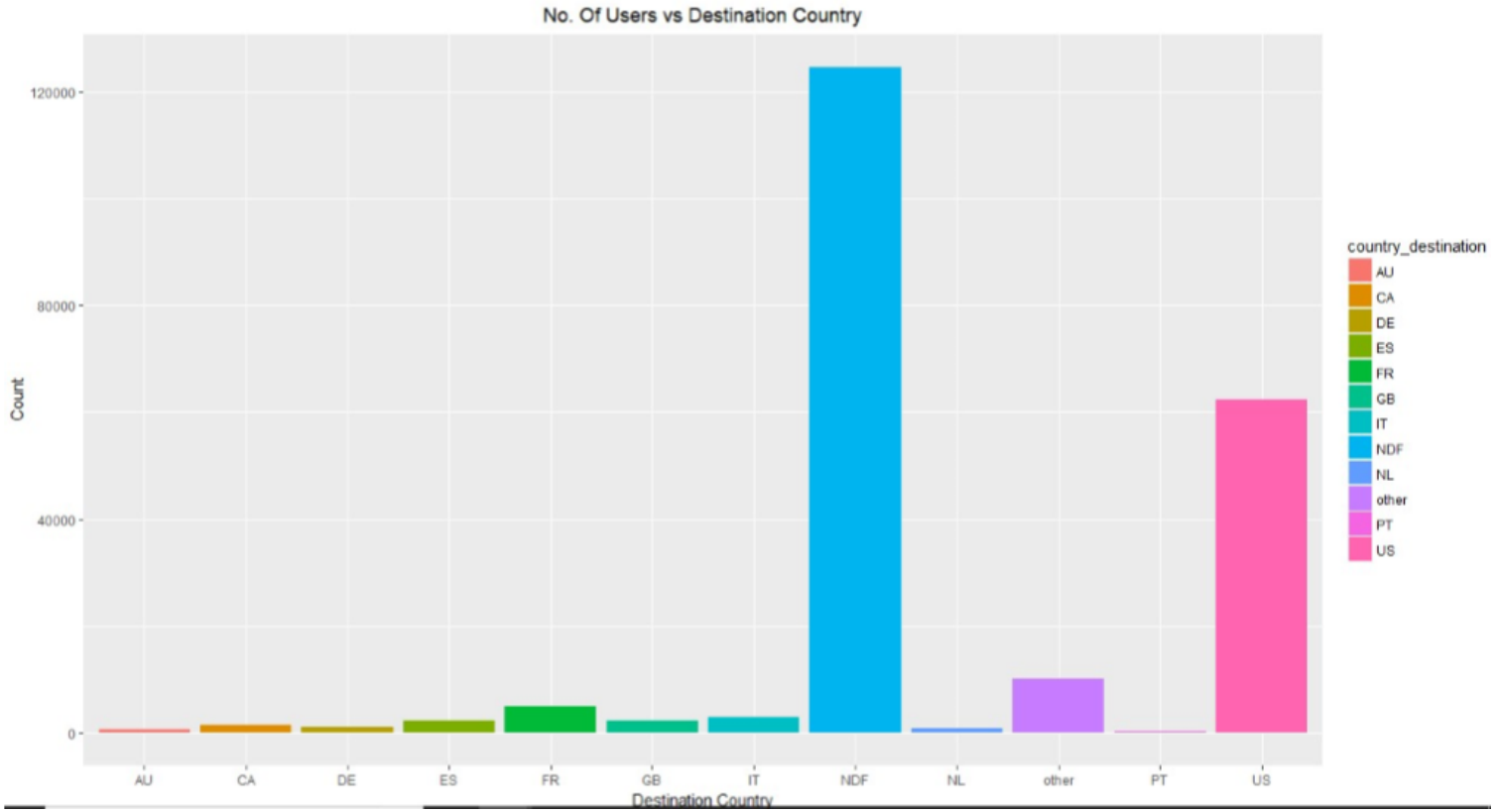
PURDUE UNIVERSITY

- Mapping number of bookings by destination country after removing users whose destination was US.



PURDUE UNIVERSITY

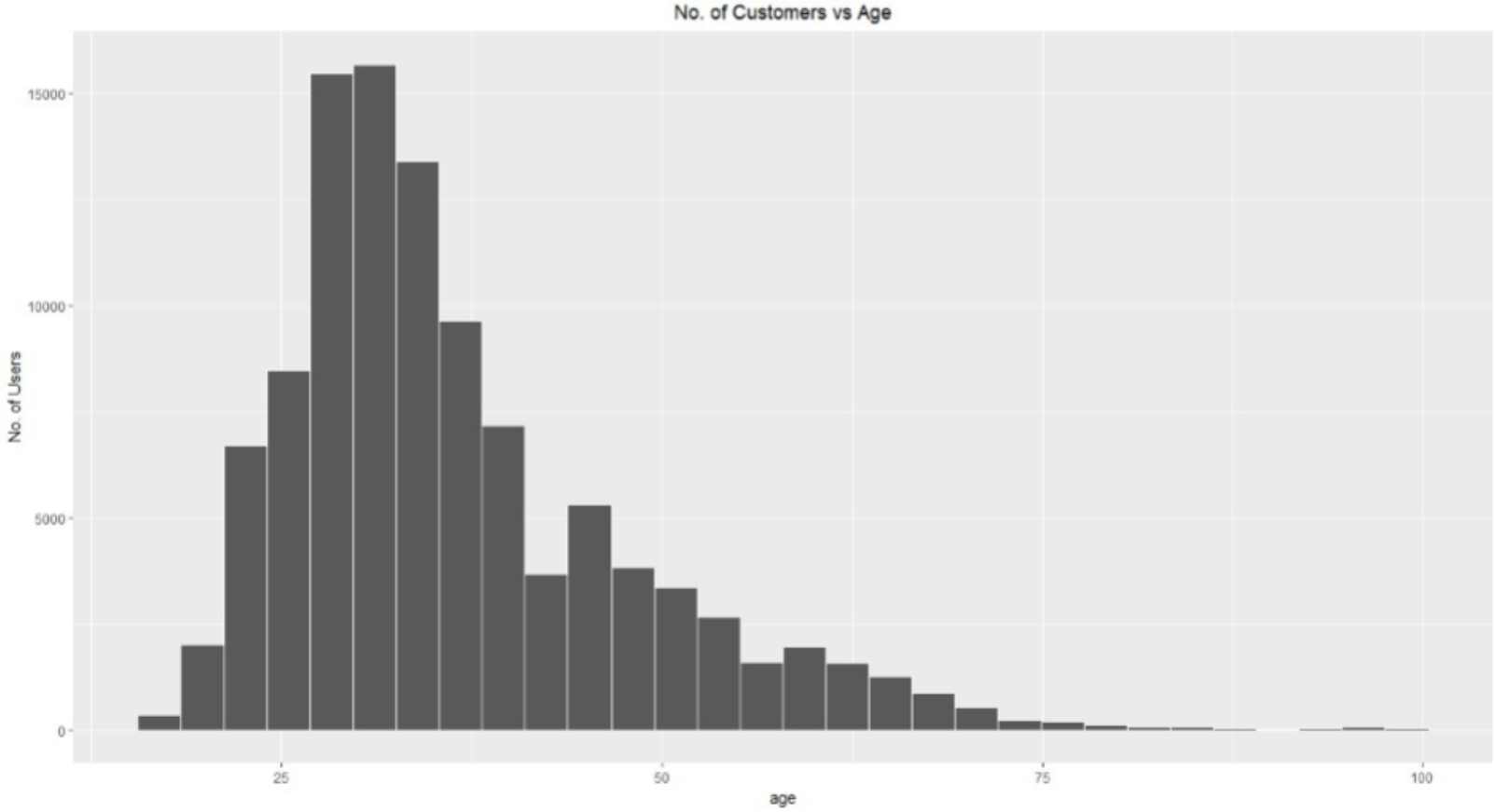
- Data distribution is skewed because maximum users are not booking



PURDUE UNIVERSITY

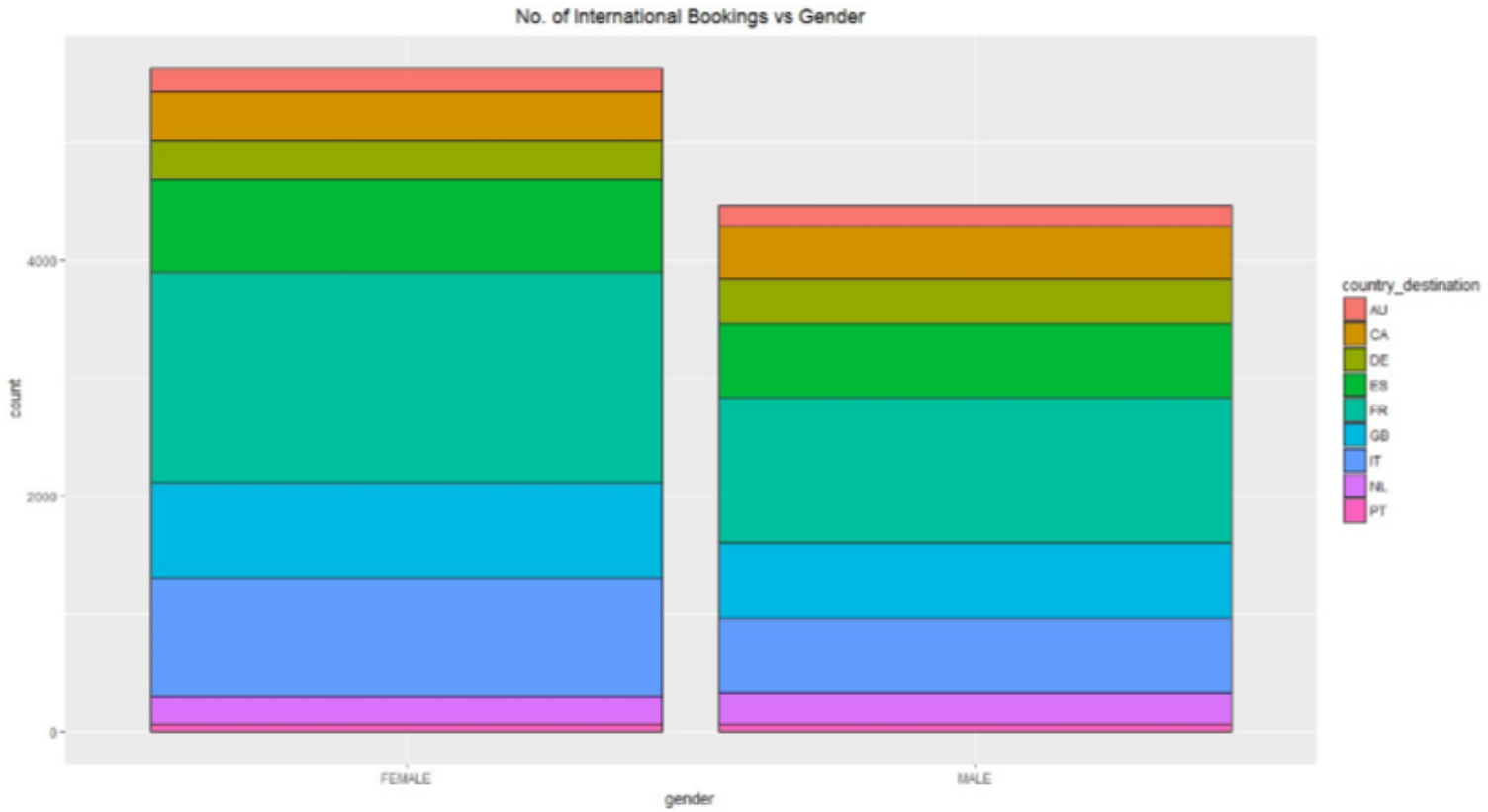
FEATURE SELECTION

Effect of Age on Bookings



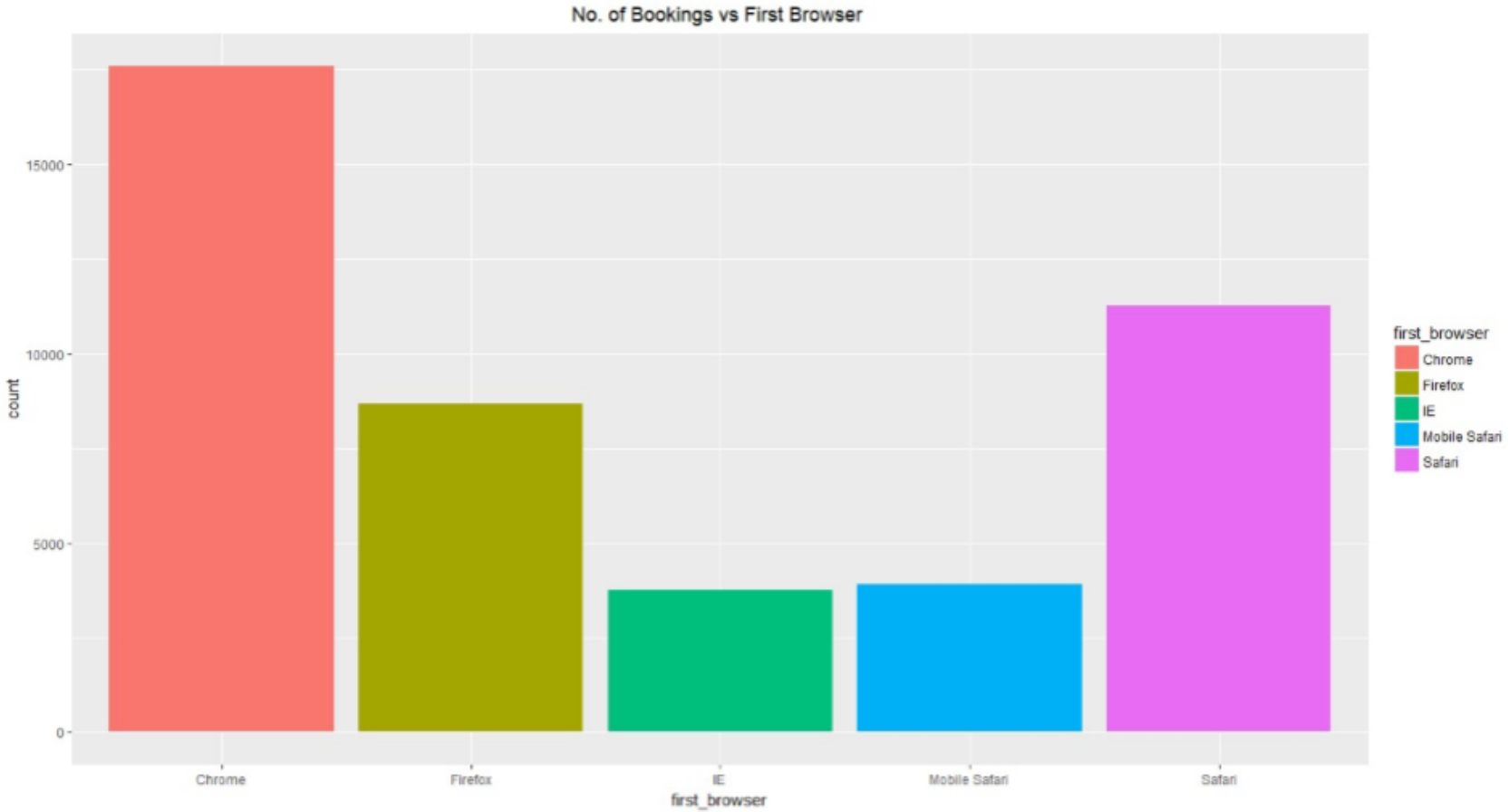
PURDUE UNIVERSITY

Effect of Gender on Bookings



PURDUE UNIVERSITY

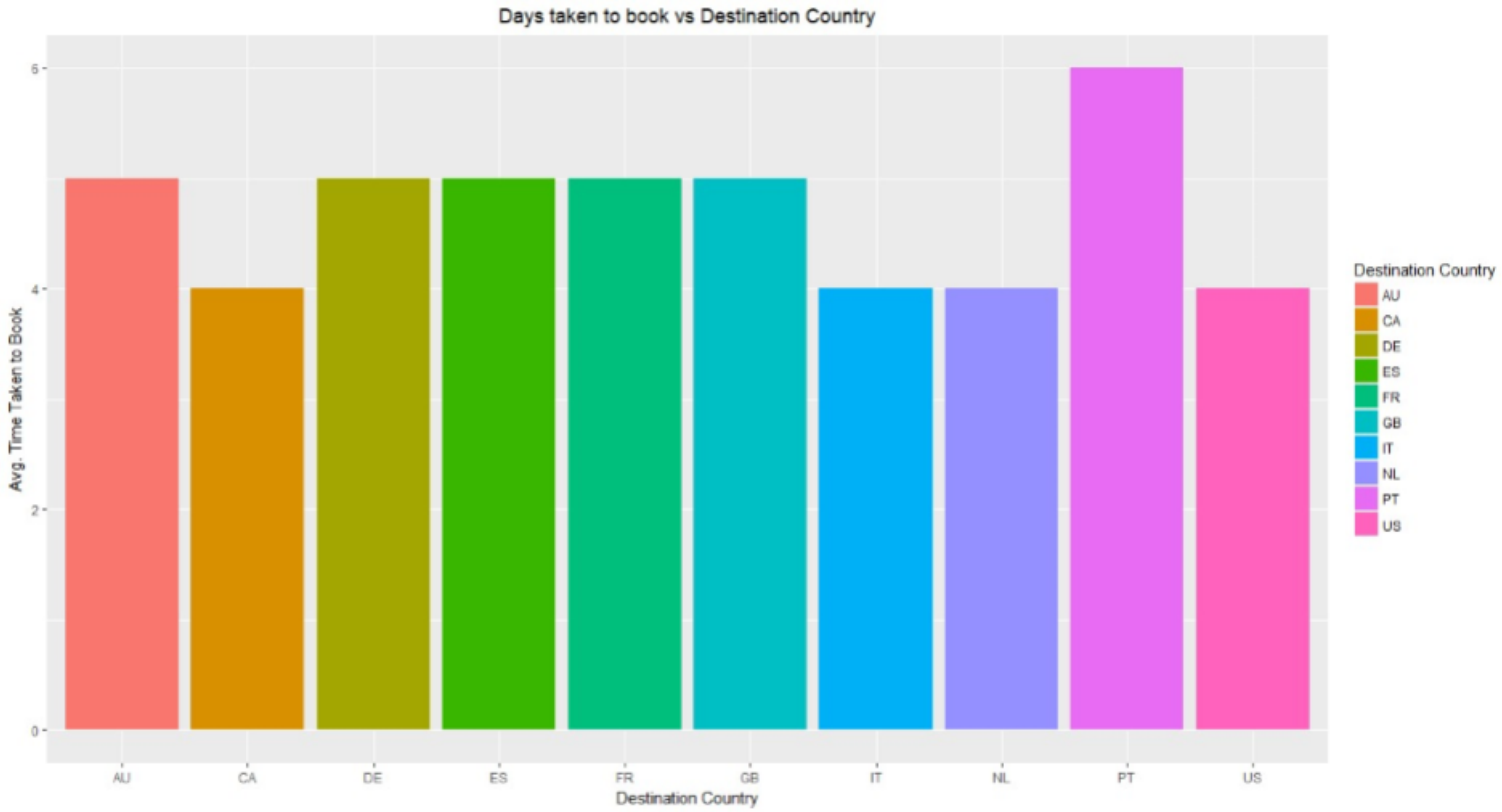
Effect of Web Browser on Bookings



PURDUE UNIVERSITY

DATA EXPLORATION

Extracted Feature-Number of days taken to book



PURDUE UNIVERSITY

METHODOLOGY AND MODEL BUILDING

Intuition

- Binary Classification: 90% of the target variable consisted of NDF and US, hence binary classification was implemented.
- Multiclass Classification: Removing the NDF's, multiclass classification to predict the destination country.

PURDUE UNIVERSITY

Models implemented for Binary Classification:

- Logistic Regression
- Random Forest
- CART
- KNN (K-Nearest Neighbors)
- SVM
- Neural Net
- Naive Bayes

Models implemented for Multiclass Classification:

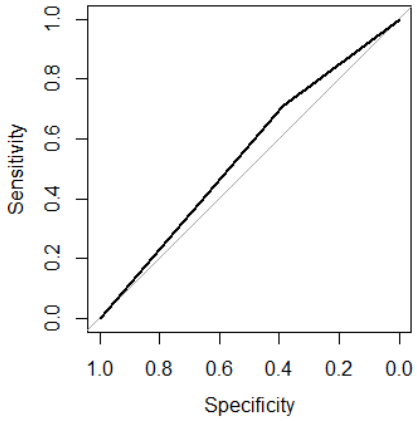
- KNN (K-Nearest Neighbors)
- Random Forest
- XGBoost

PURDUE UNIVERSITY

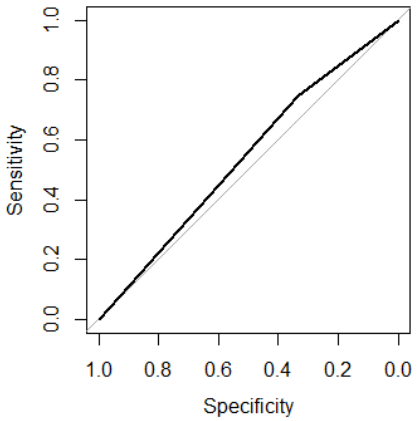
Results of Binary Classification

	Logistic Regression	Random Forest	CART	KNN	SVM	Neural Net	Naive Bayes
Accuracy	0.5559	0.5622	0.5403	0.9834	0.5761	0.5609	0.5488
Sensitivity	0.3913	0.4231	0.1761	0.9827	0.1569	0.4642	0.3361
Kappa value	0.1005	0.1148	0.0543	0.9668	0.1344	0.1132	0.0821
AUC	0.5497	0.5569	0.5264	0.9834	0.5677	0.5876	0.5407

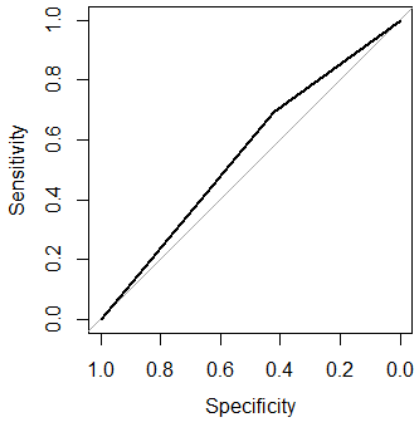
PURDUE UNIVERSITY



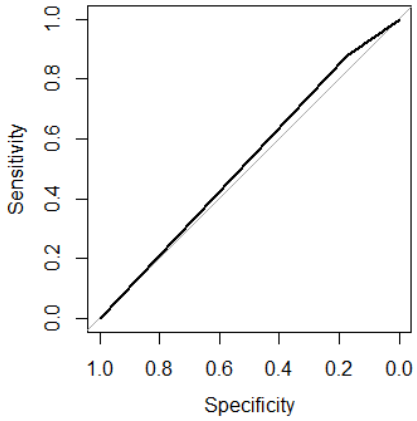
Logistic Regression



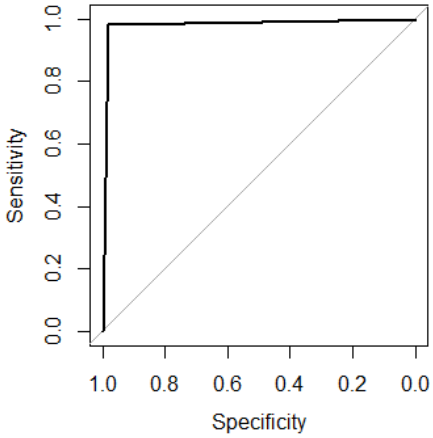
Naive Bayes



Random Forest



CART



KNN

PURDUE
UNIVERSITY

Problems faced in Multiclass Classification:

- Class imbalance.
- Class dominance (US comprising 90% observations).
- Lack of distinguishable features.
- Lack of optimal weight assignment to minority classes.
- Techniques such as SMOTE, Undersampling, and Oversampling were not effective after cross validation.

```
Confusion Matrix and Statistics

          Reference
Prediction  AU  CA  DE  ES  FR  GB  IT  NL  PT  US
        AU   0   0   0   0   0   0   0   0   0   0
        CA   0   0   0   0   0   0   0   0   0   0
        DE   0   0   0   0   0   0   0   0   0   0
        ES   0   0   0   0   0   0   0   0   0   0
        FR   0   0   0   0   0   0   0   0   0   0
        GB   0   0   0   0   0   0   0   0   0   0
        IT   0   0   0   0   0   0   0   0   0   0
        NL   0   0   0   0   0   0   0   0   0   0
        PT   0   0   0   0   0   0   0   0   0   0
        US   6  28  17  22  54  28  38   8   3 825

Overall Statistics

               Accuracy : 0.8017
                 95% CI : (0.7761, 0.8257)
    No Information Rate : 0.8017
    P-Value [Acc > NIR] : 0.5187

                  Kappa : 0
```

PURDUE UNIVERSITY

Results of KNN for Multiclass Classification

```
Confusion Matrix and Statistics

          Reference
Prediction     1     2     3     4     5     6     7     8     9    10
        1     59    33     5     2     2     0     0     0     0     0
        2    123   283    55    14     6     1     0     0     0     1
        3     14    86    82    29    18     0     0     0     0     0
        4     17    56   165   363   152    19     1     0     0     0
        5     31    62   110   448  1429   367    61     0     0     0
        6      3    16     8    30   132   316   177     0     0     0
        7      3     4     3    13    65   126   586    75     0     1
        8      0     0     0     1     2     3    29    36     0     0
        9      0     0     0     0     0     0     0     0     0     0
        10     8    29    18    50   164   138   275   227    93 27551

Overall Statistics

               Accuracy : 0.8958
                 95% CI : (0.8925, 0.899)
    No Information Rate : 0.8039
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6782
```

- 1 : Australia
- 2 : Canada
- 3 : Spain
- 3 : France
- 5 : Germany
- 6 : UK
- 7 : Italy
- 8 : Netherlands
- 9: Portugal
- 10: US

AUC=0.8668

PURDUE UNIVERSITY

CONCLUSION

- Increasing variation in the data by adding continuous variables improves model performance.
- Synthetic sampling techniques may give good in sample performance, however they have poor predictive power.
- Although computationally expensive, K-Nearest Neighbors is the best model.
- KNN performs well when majority of the predictors are categorical.
- This is because KNN captures similarity between categorical features of different target classes.

PURDUE UNIVERSITY