BA 545: Project 2

Faris Alenezy, Andre Loukrezis & Elise Vincent

Our Analytical Question...

What factors are important in predicting whether an online shopper decides to purchase a product?

Exploratory Data Analysis

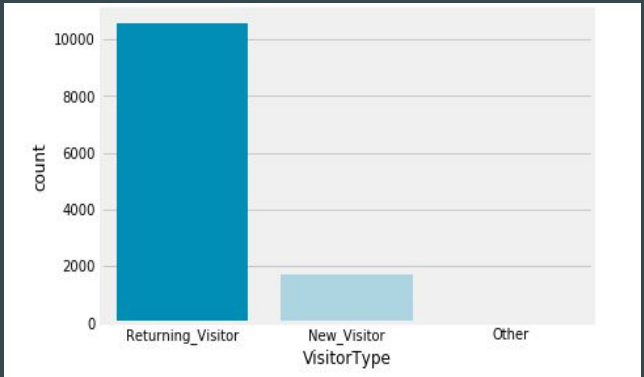
The dataset contains:

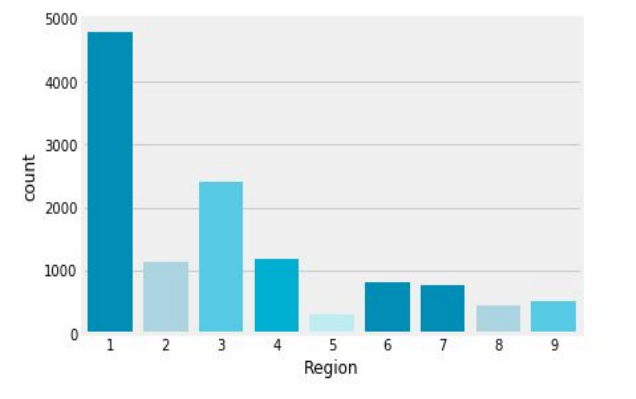
- 10 numerical
- 8 categorical features
- 1 binary target

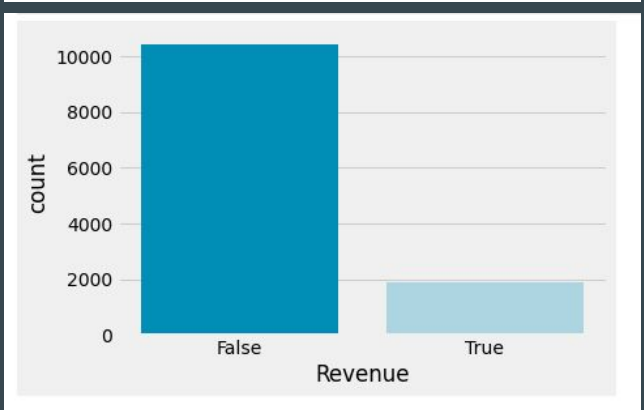
- 18 columns and 12330 rows
- 0 null or missing values

Source: *Sakar, Cemal Okan et al. "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks." Neural Computing and Applications 31 (2018): 6893-6908.*

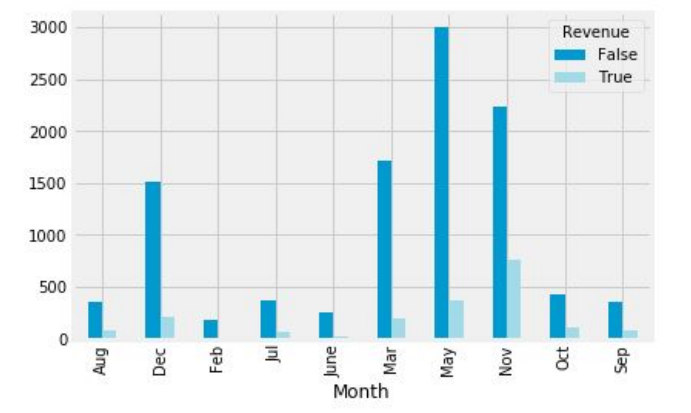
Making Sense of the Data

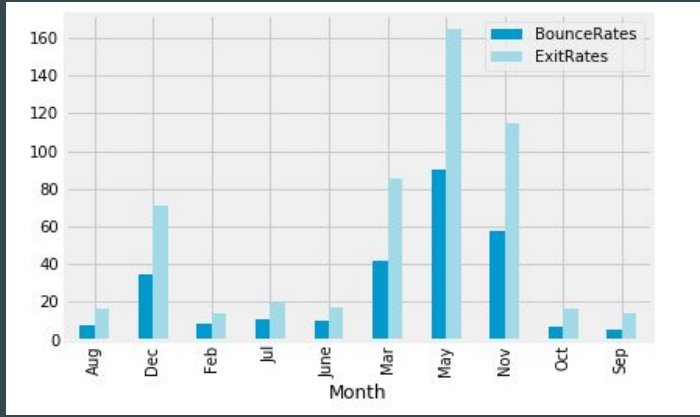


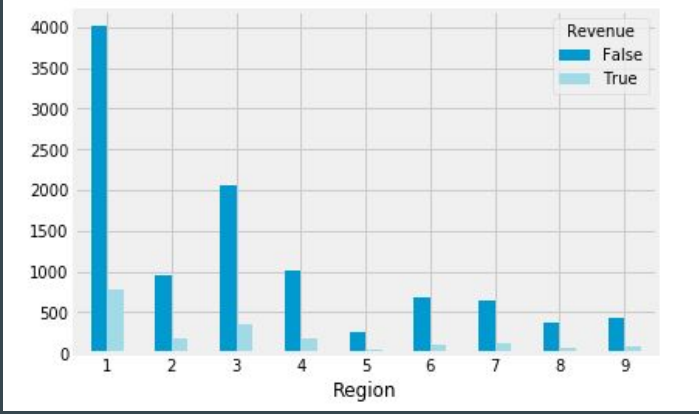


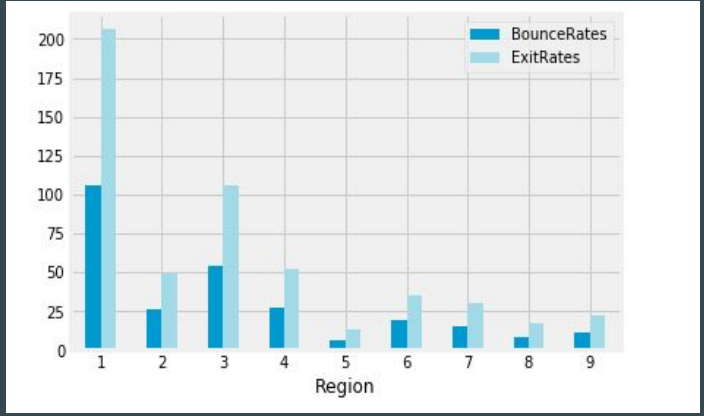


Making Sense of the Data









Correlation

- Correlation matrix highlighted several feature-pairs that were highly correlated
- FE needed to deal with high correlation

BounceRates	**ExitRates**	0.913004
ProductRelated	**ProductRelated_Duration**	0.860927
Informational	**Informational_Duration**	0.618955
Administrative	**Administrative_Duration**	0.601583

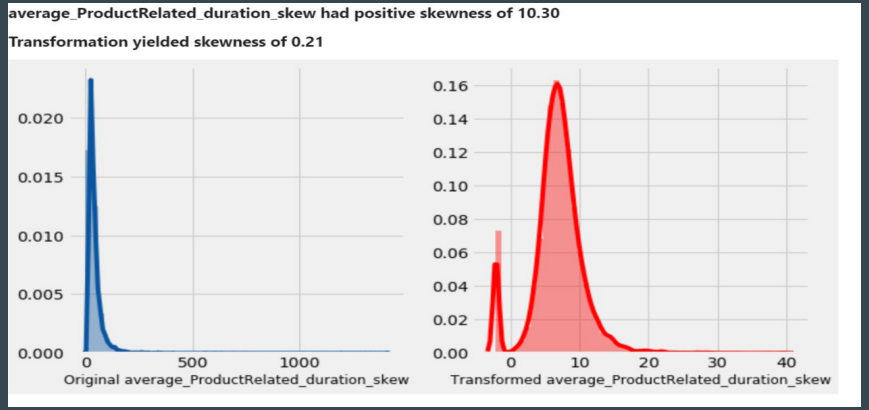
Feature Engineering

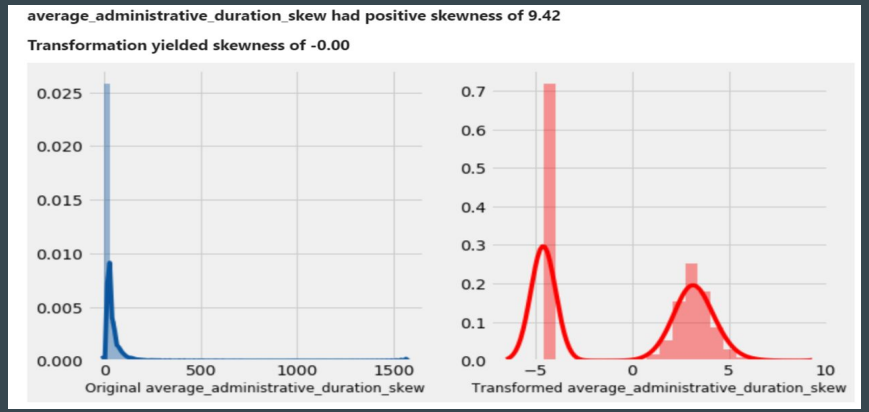
Based on our correlation matrix, we determined four pairs that required feature engineering:

	Formula
Administrative & Administrative_Duration	*Administrative_Duration/Administrative*
Informational & Informational_Duration	*Informational_Duration/Informational*
ProductRelated & ProductRelated_Duration	*ProductRelated_Duration/ ProductRelated*
BounceRates & ExitRates	$\frac{BounceRates + ExitRates}{2}$

Skewness

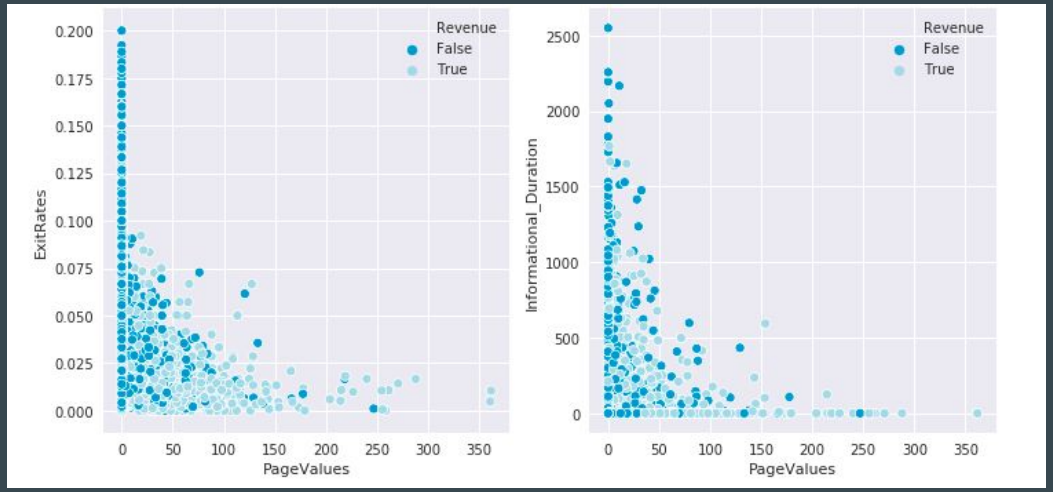
PageValues
Informational Duration
Average BounceRates/ExitRates
Average Administrative Duration
Average ProductRelated
Average informational Duration





Outliers

There are significant value to features like Exit Rates and Informational Duration that exist with “0” value for Page Values, causing the skew in our data. Because of that, we decided to not drop any outliers.



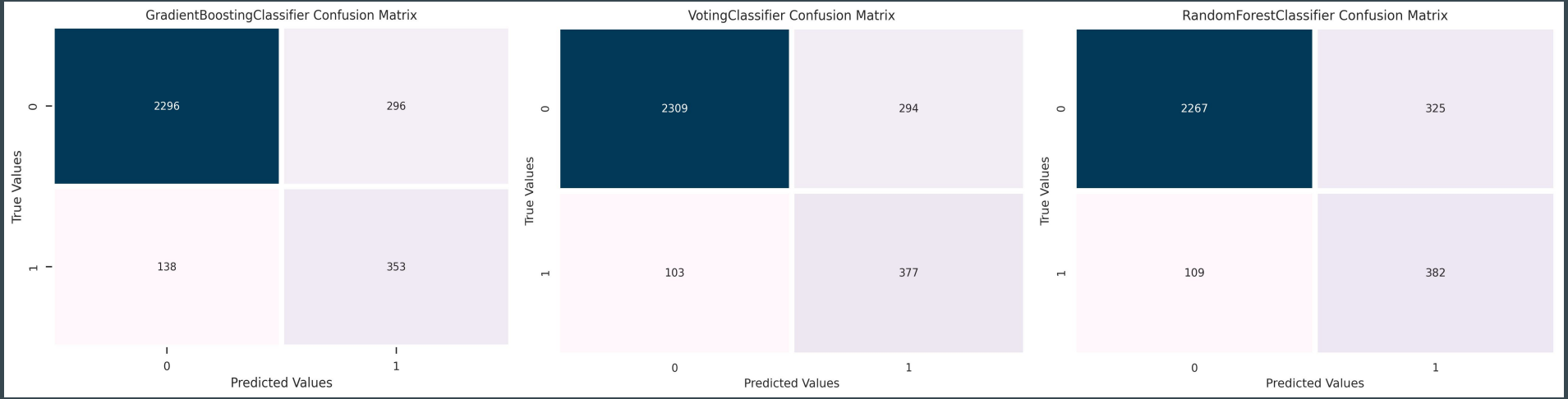
Optimizing our Models

- **Oversampling (SMOTE)**
  - Duplicating data from the minority class and adding them to the training dataset.
- **Tuning for Hyperparameters**
  - Help estimate model parameters
- **Cross Validation/K-Fold**
  -Estimate the skill of machine learning model
- **Voting Ensemble**
  - Combine the predictions from multiple machine learning algorithm
- **Feature Selection**
  - Simplifying the model and removing irrelevant features

Model Performance

	Model bias F1 score	F1 variance	Model bias AUC score	AUC variance	Model bias accuracy score	accuracy variance	Ranking
Gradient Boosting Classifier	0.9124	0.00648	0.9111	0.00665	0.9111	0.00661	1
Random Forest Classifier	0.8837	0.00861	0.8838	0.00869	0.8839	0.00868	2
Voting Classifier	0.8610	0.00903	0.8643	0.00870	0.8643	0.00880	3
Decision Tree Classifier	0.8440	0.01002	0.8505	0.00925	0.8505	0.00940	4
SVM	0.8459	0.00957	0.8512	0.00917	0.8512	0.00934	5
Naive Bayes	0.8461	0.00930	0.8476	0.00899	0.8476	0.00901	6
Logistic Regression	0.8418	0.00936	0.8441	0.00904	0.8441	0.00911	7

Model Performance



Final Model Selected

- Competition was fierce, but Gradient Boosting Classifier
- Hyperparameters:

```
GradientBoostingClassifier(learning_rate=0.05,n_estimators=1250,
max_depth=4, min_samples_split=10,min_samples_leaf=1,
subsample=0.8 , random_state=123)
```

- Slight differences using KFold allowed GBC to pull ahead of RF and VC
  - Consistent results

Important Features

	Rank
Page Values	1
Season: Fall	2
Average_BounceRates/ExitRates	3
Average Administrative Duration	4
Average Product Related	5
Average Informational Duration	6
Visitor Type: New Visitor	7
Region: 1	8
Region: 3	9
Weekend	10

Coming back to our Analytical Question...

What factors are important in predicting whether an online shopper decides to purchase a product?

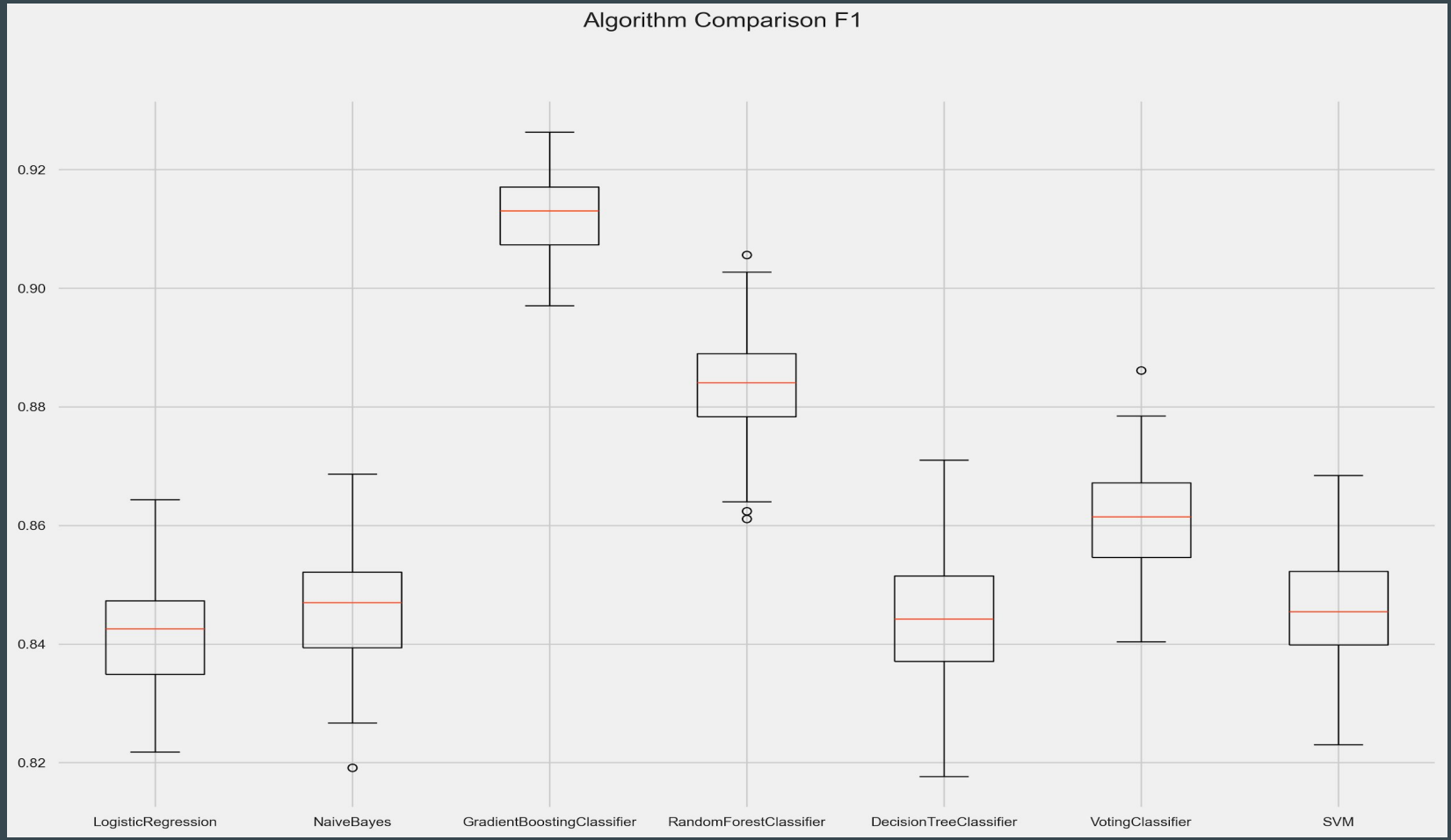
Recommendation to Increase Revenues

- Recommendation:
  - Take advantage of the administrative and information pages
    - **Add banner ads** and pop-ups for special promotions
    - Especially **target new users**
  - Test campaigns around non Fall months to increase sales in lesser periods
    - Reduced or **bundled pricing offers**, A/B test for effectiveness
  - **Increase initiatives in Regions 1 & 3**
    - Most site views and missed revenue opportunities in these regions

Potential Next Steps

- Obtain data with specific product information
  - Product type, number of units purchased, other products purchased by same customer, unit price
- Find trends of buyers' purchasing power and products similarly purchased with one another
- Create promotional ads recommending products the user may also like to purchase based on trend analysis
- Re run initial models with new, more detailed data to see if there are any significant changes to results

Questions?


Algorithm Comparison F1
0.92
0.90
0.88
0.86
0.84
0.82
LogisticRegression
NaiveBayes
GradientBoostingClassifier
RandomForestClassifier
DecisionTreeClassifier
VotingClassifier
SVM

Feature name	Feature description	Data Type
Administrative	Number of pages visited by the visitor about account management	Continuous/Float
Administrative duration	Total amount of time (in seconds) spent by the visitor on account management related pages	Continuous/Float
Informational	Number of pages visited by the visitor about Web site, communication and address information of the shopping site	Continuous/Integer
Informational duration	Total amount of time (in seconds) spent by the visitor on informational pages	Continuous/Float
Product related	Number of pages visited by visitor about product related pages	Continuous/Integer
Product related duration	Total amount of time (in seconds) spent by the visitor on product related pages	Continuous/Float
Bounce rate	Average bounce rate value of the pages visited by the visitor	Continuous/Float
Exit rate	Average exit rate value of the pages visited by the visitor	Continuous/Float
Page value	Average page value of the pages visited by the visitor	Continuous/Float
Special day	Closeness of the site visiting time to a special day	Continuous/Float
OperatingSystems	Operating system of the visitor	Continuous/Float
Browser	Browser of the visitor	Categorical/Integer
Region	Geographic region from which the session has been started by the visitor	Categorical/Integer
TrafficType	Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct)	Categorical/Integer
VisitorType	Visitor type as ''New Visitor,'' ''Returning Visitor,'' and ''Other''	Categorical/Integer
Weekend	Boolean value indicating whether the date of the visit is weekend	Binary/Boolean
Month	Month value of the visit date	Date/Text
Revenue	Class label indicating whether the visit has been finalized with a transaction	Binary/Boolean

https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset