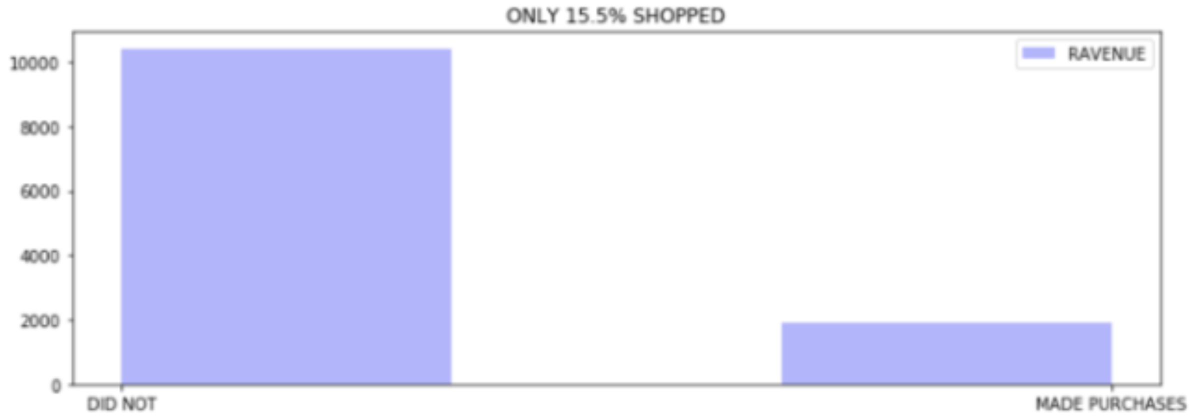
PREDICTING REVENUE BASED ON ONLINE SHOPPERS INTENTION

INTRODUCTION

This project analyzed online shoppers’ intention dataset stored on UCI Machine Learning Repository. Data set consists of more than 12,000 instances and 18 attributes that include customers’ online shopping duration data at various category including other numerical and categorical features to predict whether a customer will be ended up with shopping (revenue True) or not (revenue False).

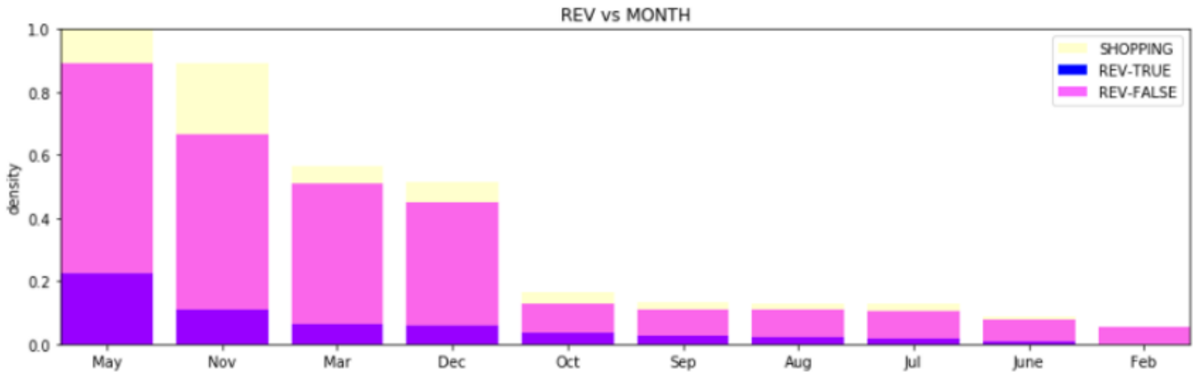
DATA VISUALIZATION

Data set consists of 15% positive class sample, i.e.; 85% did not ended up in shopping.

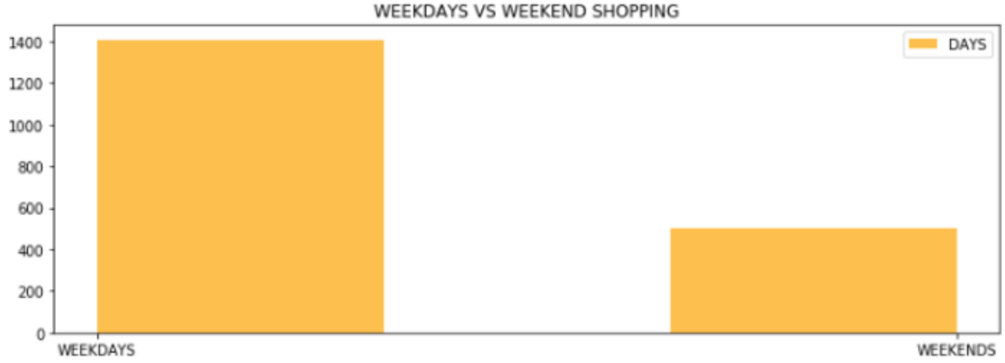


Among the eighteen attributes as listed here ['Administrative', 'Administrative_Duration', 'Informational','Informational_Duration', 'ProductRelated', 'ProductRelated_Duration', 'BounceRates', 'ExitRates', 'PageValues', 'SpecialDay', 'Month', 'OperatingSystems', 'Browser', 'Region', 'TrafficType', 'VisitorType', 'Weekend', 'Revenue'], ten of which are numerical and eight categorical.

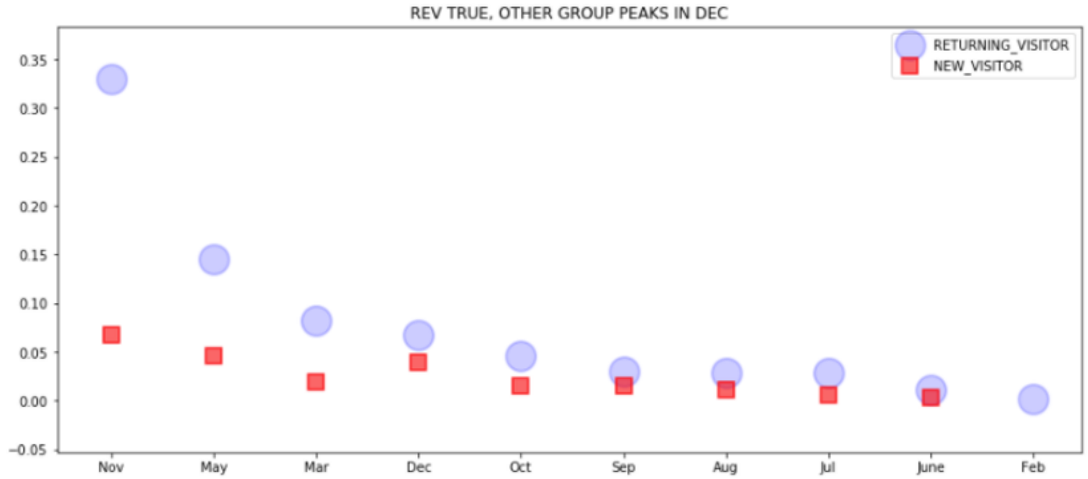
Data visualization shows that shoppers prefer to buy more in the month of May, Nov and Mar.



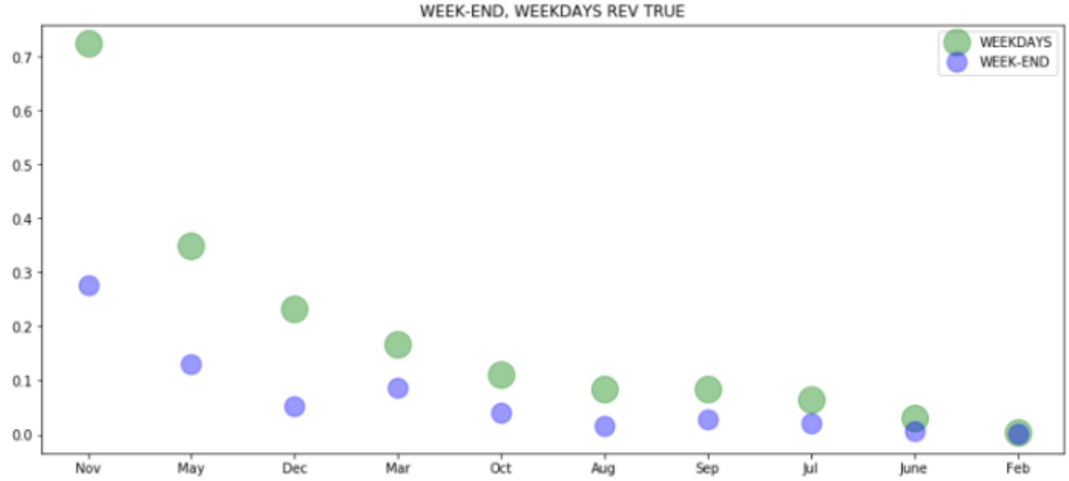
The ratio of week-day vs week-end revenue is roughly 4:1 as shown here



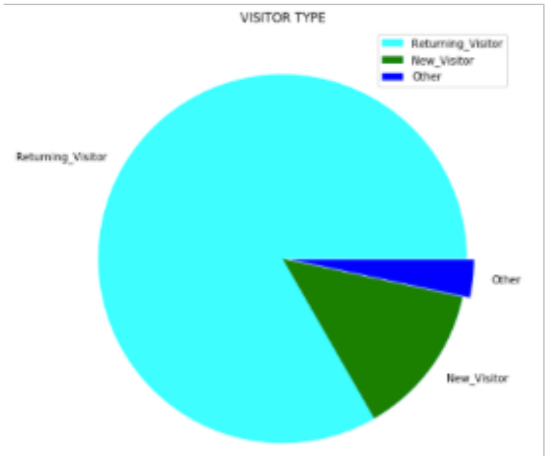
Analysis shows that the returning visitors prefers to shop more mainly in the month of Nov, May, March and Dec; however, the new visitors shop pretty much all the year round.

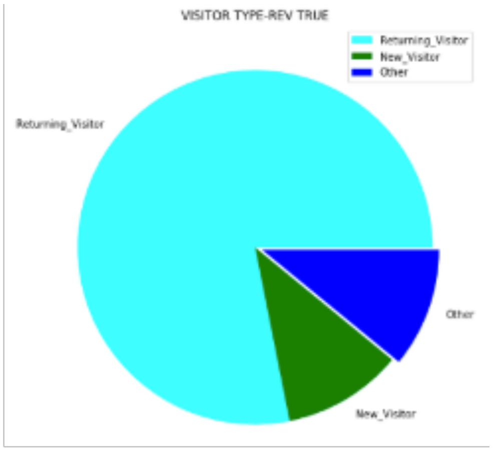


We also found that weekdays and weekends shopping vary month to month is a similar trend.



In terms of visitor types 'returning-visitor' is the largest group irrespective of whether they ended up in shopping or not as visualized in pie-charts here.



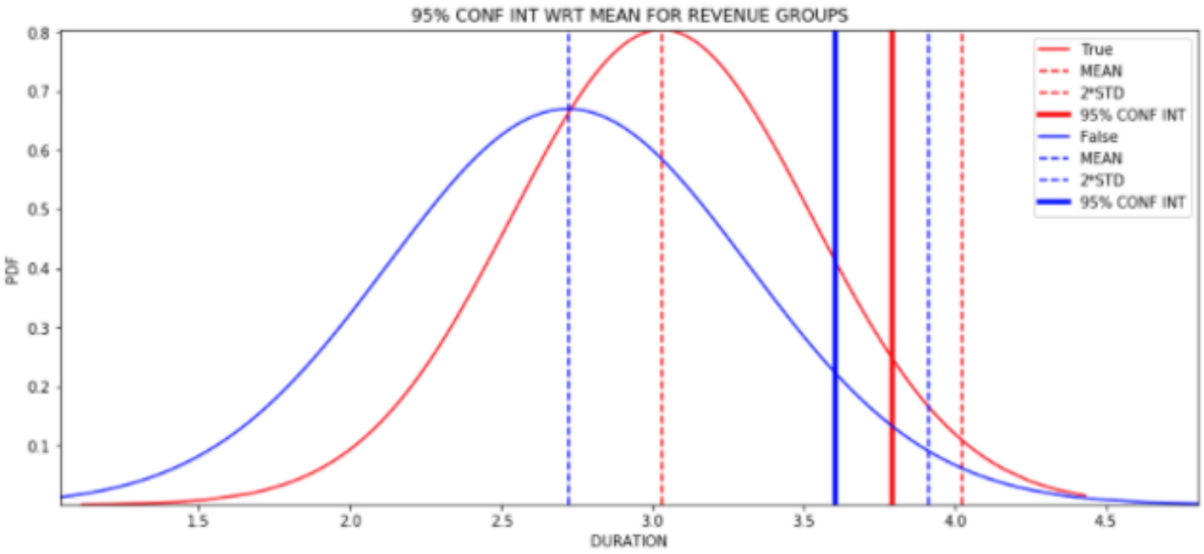


STATISTICAL ANALYSIS

While analyzing the shoppers duration in different category such as 'Administrative_Duration', 'Informational_Duration', 'ProductRelated_Duration', we found 95% confidence interval with respect to the mean within 2 STD and this holds pretty much for overall shoppers or shoppers categorized in terms of Rev-True or Rev-False as shown here, though those groups are not statistically identical in terms hypothesis test.

We also analyzed the 95% confidence interval of ‘ProductRelated_Duration’ with respect to the mean of two groups (Rev-True and Rev-False) and found shoppers who ended up in shopping spend more time online to view the product (the red curve peaks higher than that of the blue ones), however for both cases the confidence interval is close to 2 times the STD.

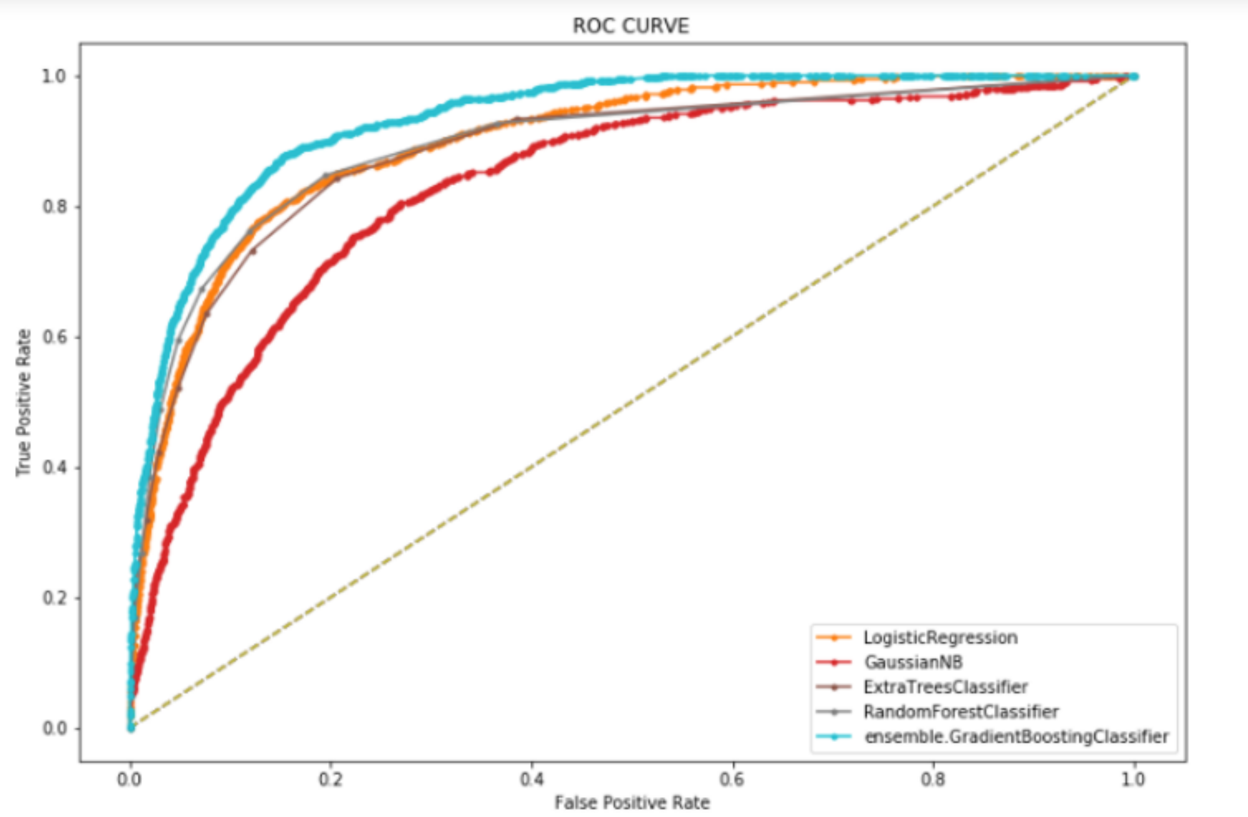
Hypothesis testing also shows Rev-True and Rev-False group are not identical.



ML PREDICTIONS

High performing Machine Learning (ML) model was implemented based on various ML algorithms such as Random Forest Classifier, Extra Trees Classifier, Logistic Regression, Ensemble Gradient Boosting Classifier, GaussianNB to predict whether the revenue is true or false for a customer based on customers online shopping duration, shopping time preferences and other attributes as well. Prediction by various ML algorithms were validated by accuracy score, ROC score, precision, recall, f1-score, confusion matrix and ROC curve.

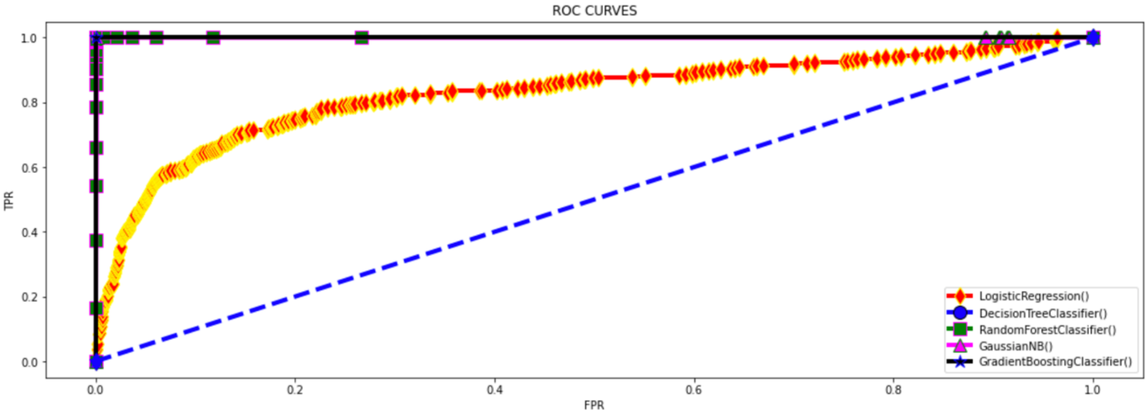
Via cross-validation, we found Gradient Boosting Classifier predicts with highest ROC score. Based on statistical analysis and ML prediction, certain offers can be made to customer to increase the revenue.



RESULTS

Analyzed shoppers' behavior and intention in terms of various attributes such as online time-spent, time preferences of the week and months of the year, along with visualizations. Also we found the largest group in terms of visitor type who did shopping is returning visitor and they mainly shop more in preferred months such as Nov, May, Mar, Dec. However, the new visitors shop pretty much all the year round. We also found the weekdays and weekend shopping vs month show similar trends. Shoppers who ended up in buying spent more time in viewing the product online. 95% confidence interval of product related duration with respect to the mean is close to 2 times the STD.

Most recent work based on rigorous data-scaling can predict with exceptional metrics score with several ML models such as decision tree classifier, random forest classifier, Gaussian naive Bayes, Gradient boosting classifier, etc.



We applied high performing Machine Learning Model to predict whether a shopper would ended up with revenue positive or negative based on various attributes such as online time-spent, shopping time preferences and other attributes as well. Based on the above prediction online shoppers can be grouped into cluster so that the companies can offer coupons to certain group of customers to increase their revenue.