MIS 373: Expedia's Strategy

Predicting Online Customer's Behaviors to Target Interventions

---

Martha Czernuszenko, Sneha Karkala, Veena Suthendran

BACKGROUND

The Expedia Group is one of the world's most well known travel platforms, and provides some of the most trusted travel brands. Expedia is an online travel agency used mainly to book airline tickets, hotel rooms, car rents, cruises, and vacation packages. The firm is one of the leaders in the $1.6 trillion travel industry.

The site uses Amadeus and Sabre (global distribution systems) for flights and hotels and Worldspan and Pegasus for bulk rate reservations. Expedia accommodates individuals looking to travel with friends or alone, families booking activities together, and large corporations planning company wide retreats. Companies which provide/plan tours and activities can apply to Expedia to form a commercial agreement. These third parties are featured on the Expedia.com for different consumers to check out. Part of the money from a purchased tour is given to Expedia.com for featuring the third party on their website. Consumers pay through Expedia's integrated payment platform for these third party activity options. After the payment Expedia distributes the money to the tour planning companies.

INTRODUCTION TO PROBLEM AND BACKGROUND

Though Expedia makes up a large component of the travel experience for millions of consumers every year, the firm has several competitors offering a similar consumer experience. Priceline, Hotwire, Orbits, Momondo, and Kayak offer very similar platforms with vacation packages, flight deals, and hotel reservation options. Consumers may browse these sites for deals, determine good hotels or airlines, read reviews, or actually make purchases. In order to compete with these other sites, Expedia offers discounts to random customers to incentivize purchases made directly through the side. However, the firm has a hard cutoff of 5% on the amount of customers the firm can provide discounts to. The firm wants to better understand their consumers to more effectively offer discounts to lock down purchases.

Three main objectives from Expedia's management team for our analyst team:

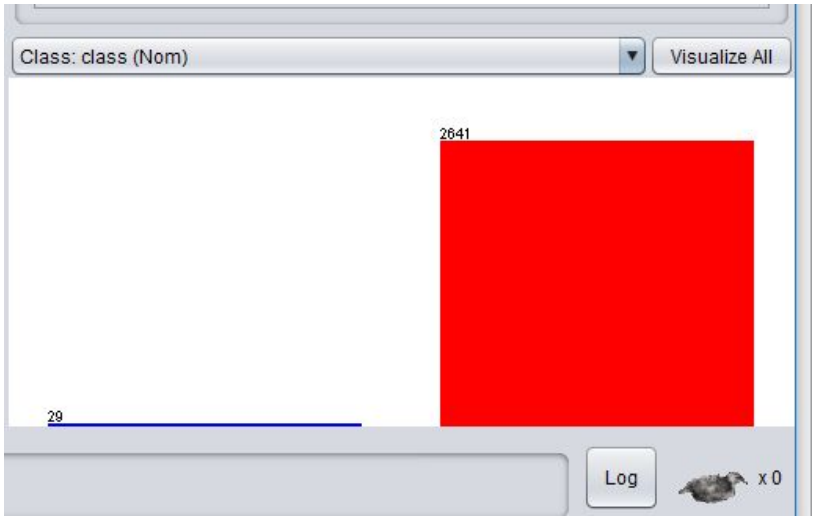
(1) Develop a model in which a consumer's decision to purchase can be predicted with the appropriate features and accuracy rate during a given online session
(2) Propose a solution which identifies shoppers and offers them coupons to increase purchases in Expedia (with currently 5% of the shoppers getting discounts)
(3) Understand whether the predictors obtained from a third party are valuable for the strategy we proposed

DATA SET EXPLORATION

The data we were provided is from Expedia.com. Each customer's session on the website corresponds to one row in the data. The data provides demographic and behavioral information about the user or consumer. The target variable, the one we are interested in predicting, is whether or not the customer made a purchase. 50% of the purchases are not completed at Expedia.com.

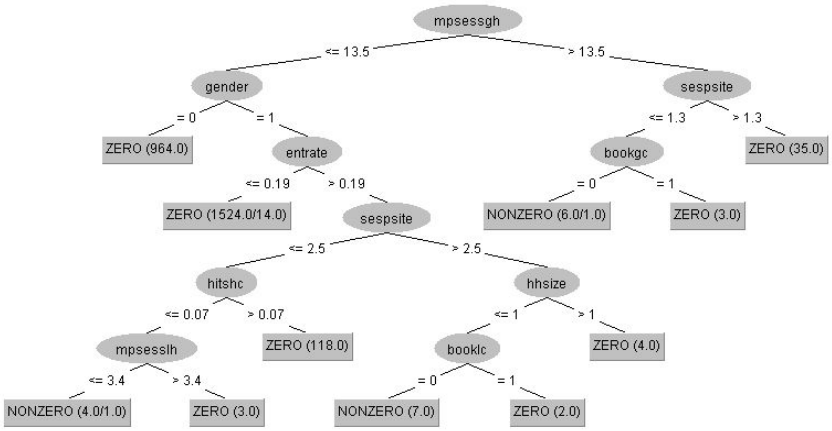
A preliminary look at a our data revealed a significant fact about our dataset: our dataset is highly unbalanced. With only 29 customers having purchased during a session, there are very few successful customers for our models to predict the behaviors of future customers on.

Exhibit 1: Graph of All Customer Classifications (1 or 0)



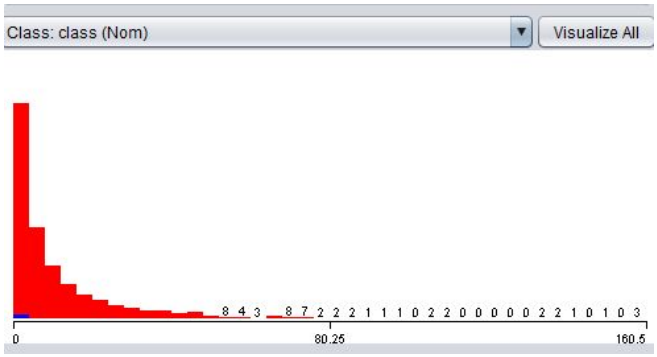
To begin our data exploration, we started by developing a classification tree.

Exhibit 2: Classification Tree



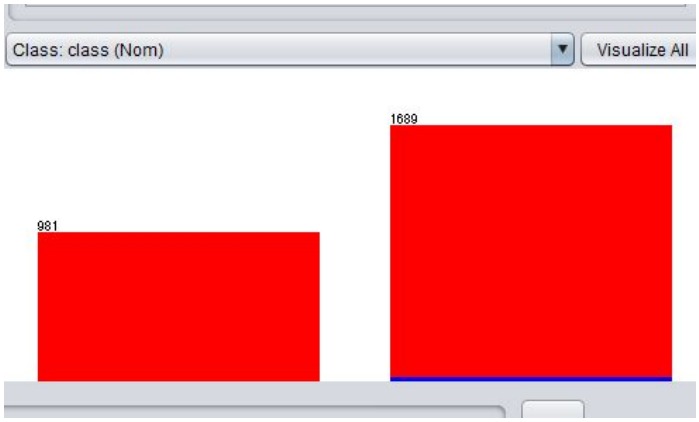
As seen in Exhibit 2, mpsessgh is the first split in the tree with mpsessgh <=13.5 or mpessgh > 13.5. This indicates that mpsessgh is the attribute in the dataset that has the greatest effect reducing the entropy of the entire data set. Based on the division of the relevant class in each attribute chart, other predictors that seem to have a significant impact on the classification of a customer are sespite, gender, entrate, and bookgc.

Exhibit 3: Histograms of various attributes in dataset
Minutelc: Time Spent up to this point in the set



Sessgh: Total no.of sessions visits of all sites so far

Gender: 1- Male, 0 -Female



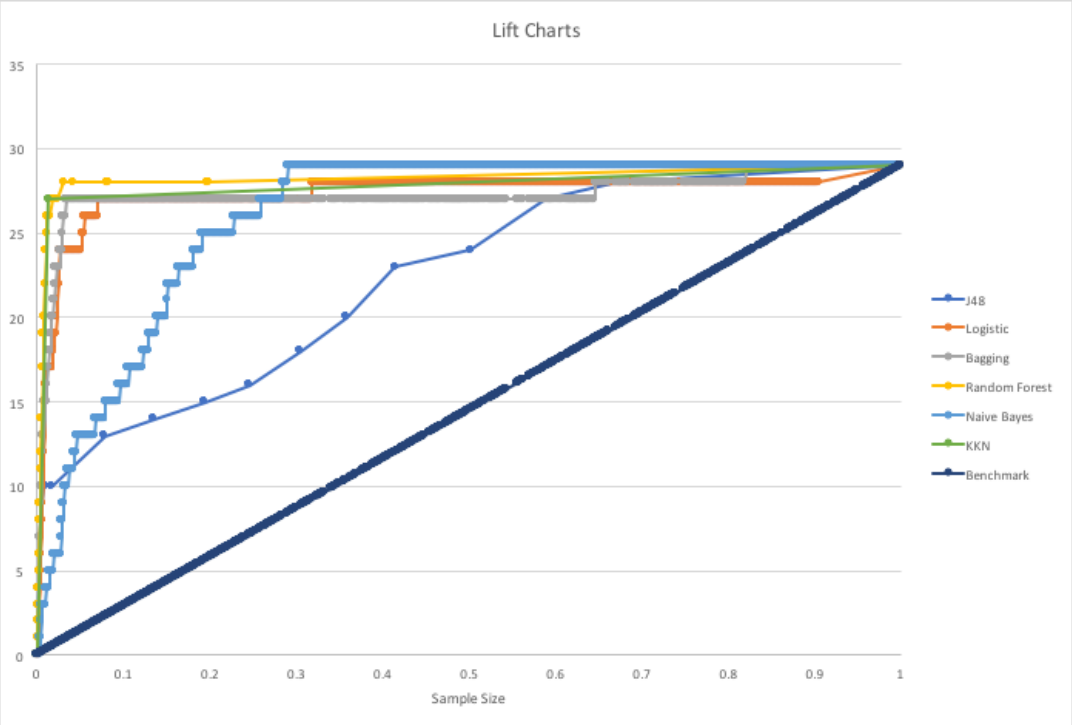
It is also particular to note, as evident in Exhibit 3, our dataset is heavily unbalanced when it comes to gender: there are only 29 customers who purchased and they are all men. Other predictors that might have predictive value are gender, sessgh, and minutelc based on the patterns observed above.

DATA ANALYSIS

Lift Chart

After exploring potential attributes of importance within our data set, we took to WEKA to determine the best model to use for Expedia to predict visitor behavior. With several potential models to choose from, we analyzed which model would yield the best results for Expedia given the company's constraints and desired deliverables. In total, we considered six differents models: J48 Classification Tree, Logistic Model, Bagging Model, Random Forest, Naive Bayes and K-Nearest Neighbors (10 Neighbors). After running each model, we took the sample size and lift provided by our WEKA analysis and plotted all six models against a benchmark, or true number of customers that made a purchase during the session (see Exhibit 4 below). The benchmark for the provided data set was 29 customers that made a purchase out of 2670 customers, or 1.086% of customers.

Exhibit 4: Lift Chart of Potential Models



Based on our lift chart, it appears that the Bagging, Logistic, Random Forest, and K Nearest Neighbors follow similar paths significant above the benchmark. Random Forest has the highest lift till .3 and then Naive Bayes has the highest lift. Expedia might want to consider using different models for different sample sizes. However, given the limitation a campaign budget that offer discounts to up to 5% of shoppers, the models provide the following lift estimates:

Exhibit 5: Lift Estimates of Each Model @ Point Reaching 5% of Shoppers

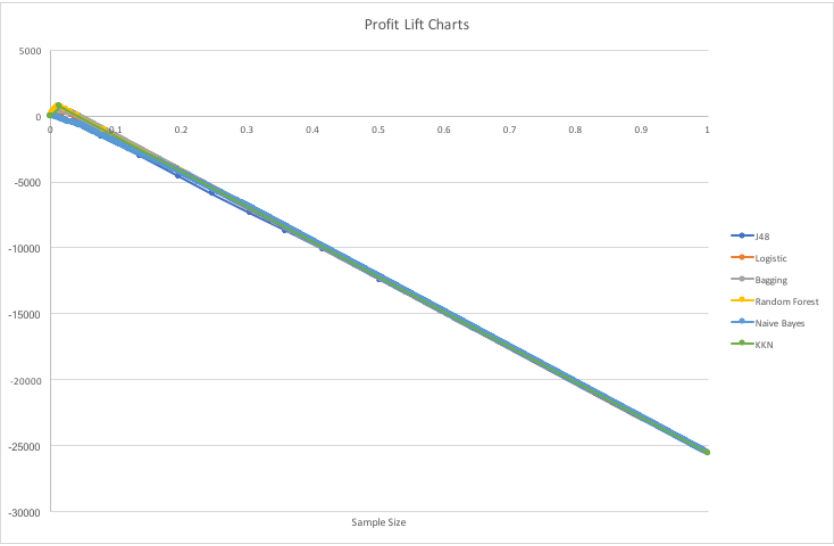
Model	**Lift @ 0.05**	**Model**	**Lift @ 0.05**
J48	~8	Random Forest	28
Logistic	24	Naive Bayes	13
Bagging	27	K Nearest Neighbors	~27
Benchmark	1.45		

Thus, based on the budget limitations of Expedia, the Random Forest Model offers the highest potential lift (28) and should be the model used for this campaign. If Expedia were to use our proposed model, we would capture 28 customers or 20.9% of the target (5% of shoppers). Our model captured 28 out of the 29 buyers in the population, which is very effective.

Profit Chart

Thus far, the lift chart analysis suggests that the Random Forest model offers the most viable model for Expedia. However, it is also important to consider the profit lift charts of the six models considered in making this decision. Based on assumptions provided by Expedia we created a cost matrix (see Exhibit 7) to use to create a profit chart for each model: "a purchase yields a $40 premium to Expedia, and the discount is $10." The graphs of each model's potential profit can be see in Exhibit 4 below.

Exhibit 6: Profit Chart for Potential Models



Zoomed In Profit Chart

Exhibit 7: Cost Matrix

Predicted (A)	*Predicted (B)*	
30	0.0	*Actual (A)*
-10	0.0	*Actual (B)*

Exhibit 8: Maximum Benefit

Model	**Max. Benefit**	**Model**	**Max. Benefit**
J48	Benefit: 230 Random: -162.61 Gain: 392.61	Random Forest	Benefit: 730 Random: -295.52 Gain: 1026.52
Logistic	Benefit: 390 Random: -277.39 Gain: 667.39	Naive Bayes	Benefit: 0 Random: 0 Gain: 0
Bagging	Benefit: 370 Random: -411.32 Gain: 781.32	K Nearest Neighbors	Benefit: 730 Random: -334.8 Gain: 1064.8

Based on the zoomed in profit chart (Exhibit 6) and maximum benefit table (Exhibit 8), it is evident that both the Random Forest and K Nearest Neighbors model yield a maximum profit of $730. However, Random Forest is able to achieve this profit by targeting only 1.161% of the

population, with a probability of .3 or 30% (score threshold) or more. This 1.161% requires less of Expedia's budget towards customer targeting and therefore is this most optimal model to use going forward.

Feature Selection

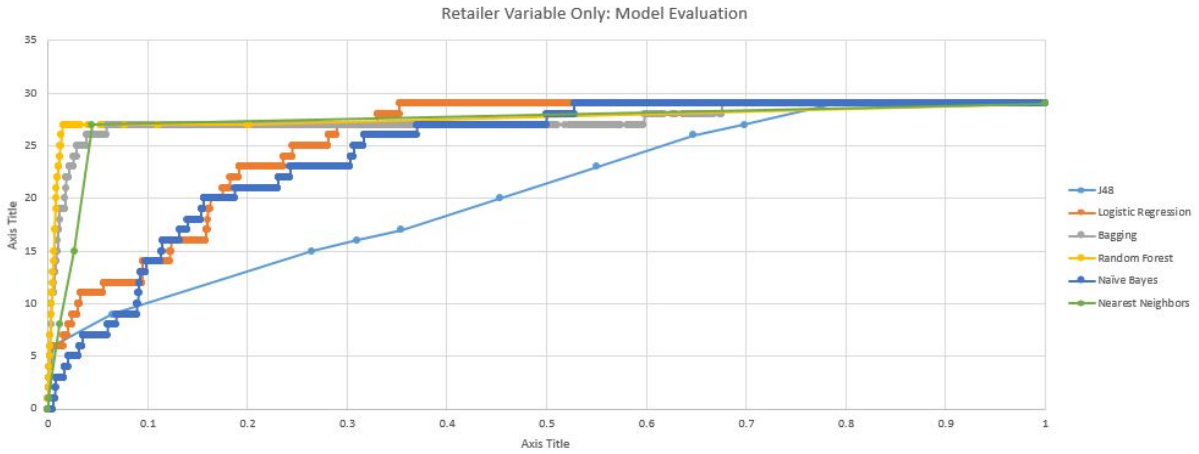
Once we determined the Random Forest Model was optimal based on the lift chart, we used the ROC area to select the features important to the prediction. We first determined the initial model ROC area, then after ranking the attributes based on information gain, we took out the predictors one by one and redid the model to see the updated ROC area.

The initial model had an ROC area of .978, and the final ROC area was .979. We kept the following variables in our model, because when we took them out the ROC area decreased: Sesslh, Exitrate, Minutelc, Minutshc, Entrate, Minutegc, Sessgh, Bookgh, Single, Booksh, Booklh, Minutelh, Hpsesslh, Awareset, Bookgc, Weekend, SEgc, Path, Edu, Sessh, SErate, Hitsch, Child, Httlc, Hhsizem, Income, Gender, Peakrate, Mpsesslh, Mpsessgh, Minutegh, and Age. This confirmed our initial hunch in our data exploration that minutelc, sessgh, and gender are critical predictors.

Data Analysis for Retailer Variables only

Furthermore, we wanted to determine which variables are the most informative. Variables 1-14 are owned by the retailer while Variables 15-38 were purchased by a vendor. Our group decided to remove the variables purchased by a vendor to determine the impact of just retailer variables. We kept 2 variables that were not included in the variable description because they are not classified as retailer or vendor, therefore had 16 variables. In total, we considered six differents models: J48 Classification Tree, Logistic Model, Bagging Model, Random Forest, Naive Bayes and K-Nearest Neighbors (10 Neighbors - because we tried using 1 and it was based on an outlier). Once we ran each model, we found the lift curve through WEKA analysis.

Exhibit 9: Lift Chart of Potential Models for Retailer Variables Only



Based on our lift chart, we can see that the Random Forest would be best for the first 5% of the our target. Later, other models such as nearest neighbors perform similarly to 30%. After, Logistic Regression dominates and then Naive Bayes. Since we are looking to target for the 5% of the target, we should use the random forest model.

Exhibit 10: Lift Estimates of Each Model @ Point Reaching 5% of Shoppers: Retailer Variables

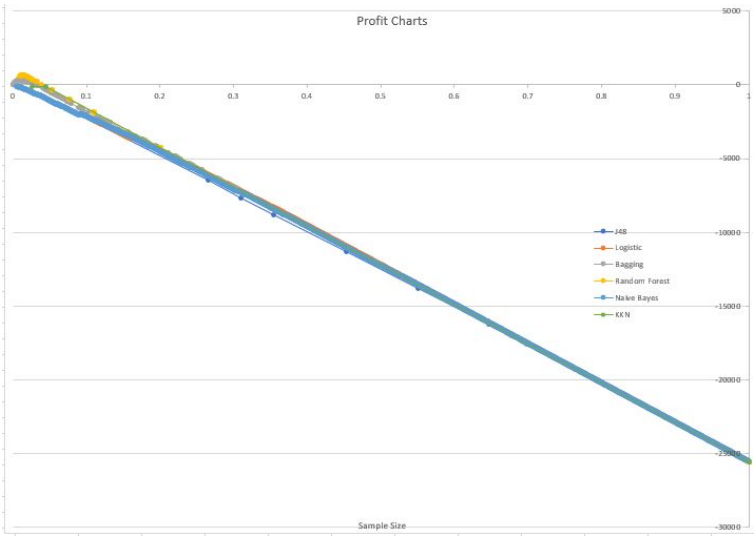
Model	**Lift @ 0.05**	**Model**	**Lift @ 0.05**
J48	~6	Random Forest	27
Logistic	11	Naive Bayes	7
Bagging	26	K Nearest Neighbors	~27
Benchmark	1.45		

Thus, based on the budget limitations of Expedia, the Random Forest Model offers the highest potential lift (27) and should be the model used for this campaign since it has a higher lift than K Nearest Neighbors. Our model captured 27 out of the 29 buyers in the population, which is really effective (20.2%). However, the dataset with all of the variables was able to capture 28 customers. All customers yield the same revenue, so Expedia would want to consider the potential costs of gaining all of the variables in order to capture this additional customer. Furthermore, due to the limitation that are dataset is unbalanced, we would want to consider if it is an outlier.

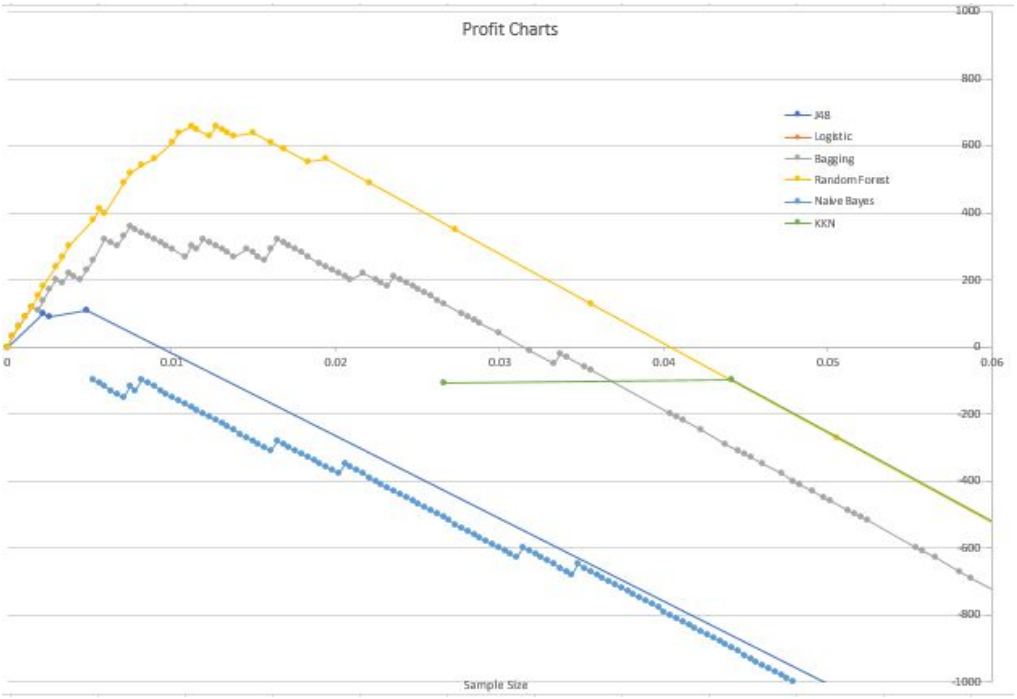
Profit Chart

In order to construct these profit charts, we used the same cost matrix as the one for all the variables. Through these profit charts, we found that the most effective model would be Random Forest. The probability threshold is .2333, and it hits about 89% of our target, and 1.237% of the population. We should offer the discount to 1.237% of the population. Furthermore, the expected increase in revenue to Expedia.com would be the gain of 985.23 and a profit (benefit) of 660. An important thing to consider is that are dataset is highly unbalanced- only 29 customers said yes. With more data, we might be able to see a more illustrative result.

Exhibit 11: Profit Chart for Potential Models - Retailer Variables Only



Zoomed in Profit Chart for Potential Models



Feature Selection for Retailer Variables

The Random Forest Model was also optimal for the variables from the retailer. We used the same method as before to determine the attributes which better the prediction ability of the our model. We first determined the initial model ROC area, then after ranking the attributes based on

information gain, we took out the predictors one by one and redid the model to see the updated ROC area.

The initial model had an ROC area of .960, and the final ROC area was .979. The initial model had an ROC area of .978, and the final ROC area was .979. We kept the following variables in our model, because when we took them out the ROC area decreased: Hpsesslh, Httlc, Gender, Mpsesslh, and Age. From our initial data exploration, we confirmed that gender is an important predictor.

RECOMMENDATION

We recommend that Expedia target customers via a Random Forest Model because this had the highest lift and benefit. Through analyzing the lift and profit lift charts of six different models using two subsets of data, one including external data and one without, we determined that the Random Forest Model is the best overall.

By using this model, we recommend targeting 1.161% of the population, significantly less than the budgeted 5% of the model, in order to generate $730 in net profit for the company per 2670 customers or $0.27 per customer. With our model, we are able to save more in our budget by not having to pay the price of capturing the other ~4%. Our data analysis indicated that a net profit of only $660 could be obtained with the internal data. By buying the external data, we are able to increase our profit to $730, so if the external data can be purchased for less than (730-660)/2670 customers = $0.048 per customer, it is worth the investment. The Random Forest model with external data will yield a revenue of $1026.52. The K Nearest Neighbors model does offer a higher potential revenue of $1,064.80 at the same profit level of $730; however, would require a higher cost as it requires targeting more customers than the Random Forest Model. Our analysis links directly to Expedia's business objective of improve the efficiency of retailer's online discount offerings by targeting specific customers who have a high probability of actually purchasing. This is in contrast to the current strategy of randomly selecting 5% of the customers. We would recommend purchasing the vendor's data attributes since it led to increased profits. However, we need to make sure that the price of the data is worth is the $50 difference in profitability.

Additionally, the ROC area of the model with external attributes and for just the retailer variables was 0.979, which means that the predictive ability for both of these Random Forest Models is the same. This ROC area was calculated after removing variables which lowered or didn't affect the predictive ability of the Random Forest Model. The model for the internal data plus the vendor data and for the model with just the internal data have the same predictive ability.

CONCLUSION

Through our analysis of lift charts (looking at who to target), profit (looking at expected increase revenue and the discount budget), and feature selection, we were able to recommend a comprehensive strategy to Expedia. As the travel market shifts from travel agents to online, it is

important to consider customer behaviors which are indicative of a future purchase. Expedia currently targets randomly at 5%, so through our analysis and recommendation, Expedia can predict online customer's behaviors more effectively to target interventions.