Project Report

A Business Case-study: Using Data to Bring Customers Home

Alefiya Naseem, Devanshi Gariba

DS5500 Applications in Data Science Fall 2019

naseem.a@husky.neu.edu

gariba.d@husky.neu.edu

I. Introduction

The game of online retail and e-commerce has changed a lot over the past few years. Offline purchase and local stores have been replaced by hosting online stores. This report explores a business case-study of one such e-commerce company: Wayfair. Wayfair is an e-commerce retailer that sells furniture and other home goods. While they sell to individual consumers, they also have a large B2B (Business-to-Business) division that sells to business customers such as interior design firms, contractors, hotels and universities.

As a data-driven company, they ensure that their B2B customers receive best-in-class service by leveraging data science models to predict customer needs and purchasing patterns. They primarily aim to reinvent the way the world shops for home and utilizes machine learning models heavily in many of their departments including marketing, sales, and operations teams to guide business decisions. Often times, business vendors bring in most profits and it becomes really important for a company to personalize efforts for them and retain them and take appropriate measures if they are not retained.

Keeping this in mind, Wayfair hosted a challenge this year for which they released their B2B customer interaction dataset that will be used here. The goal of the challenge was to build a model to predict customer behavior for Wayfair. The goal of our project is to gain insights from the business customer data like customer information, sales call records, purchase history etc. and build predictive models to work on the following problems:

1. B2B customer conversion (classification): Whether a B2B customer will purchase or not in the next 30 days
2. B2B customer expected revenue (regression): How much a B2B customer will spend in the next 30 days.

Our case-study focuses on certain research questions that help us reach towards our final goal for prediction of B2B customer conversion and the expected revenue. However, one of our main goals is to be able to focus on the interpretable and explorable approximations of Black Box machine learning frameworks. Oftentimes, we have high dimensional customer interaction datasets to work with and the goal ultimately shifts from *understanding the customer behaviour analytically to prediction problems that are mostly handled by the black-box algorithms.* Our work heavily explores multiple feature selection techniques and identification of different feature subsets. Important feature subsets can help in analytically understanding the customer behavior. Identification of actionable features associated with customer retention can mainly help businesses turn their non-retaining customers into retaining customers if we can identify the main features for them and use business strategies around those features. Some of the research questions that we used were:

1) Can we use data adaptive machine learning algorithms to predict Whether a B2B customer will purchase or not in the next 30 days.
   a. There are an overwhelming number of features in the dataset. What are the most

important features that help in the prediction of the conversion.

b. What techniques can be useful in determining the right feature-set?
c. Determine the optimal number of features in predicting whether a customer will purchase or not
d. Infer different feature sets to see if there is a logical explanation and interpretation of why some features are important over others and see if we can personalize efforts on customers that don't convert.
e. What modelling strategies best help in predicting the score?

2) What are the most appropriate methods for tidying the dataset? How to handle missing data, categorical data?
   a. How do we make sure our results are not biased?
   b. Will missing data imputation improve our prediction or yield high-performing feature-sets?
   c. Does missing data have any effect on feature
   d.
   e. subset selection?
3) Data-set is highly unbalanced. What are the appropriate sampling techniques to deal with class imbalance since there's a huge amount of missing values too?
   a. Identify and implement appropriate resampling techniques to deal with class imbalance.

4) How to develop an intuition from the dataset given the high dimensionality of the problem?

a. Are we able to produce interesting visuals demonstrating the relationship between the target variable and the associated features?
b. Perform exploratory Data Analysis

II. Dataset

The entire dataset is divided into two parts:

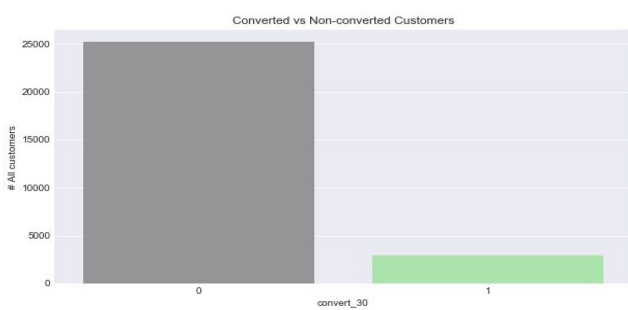
1. Training data: This data includes 181 features which can be divided into different classes depending on their functionalities. These are:
   a. Two outcome variables: This class contains the two outcome variables- convert_30 and revenue_30
   b. Customer: This class contains basic information about a customer like type, role, team, status, etc
   c. Enrollment questions: This class contains information like number of employees, number of purchases, total cost of purchases per year, etc
   d. Order: This class contains order information like number of orders, size, influence, etc over a time frame
   e. Satisfaction: This class contains customer satisfaction information over a period of time
   f. Visit: This class contains information like number of visits to the site or favorites list over a time frame
   g. Search
   h. SKU: This class stands for Stock Keeping Unit. It contains information like average price of a product viewed over a time frame, etc.
   i. Task: This class contains task information like introduction, cadence, reassignment and others over a time frame
   j. Call: This class contains call information over a time frame
   k. Email-BAM: This class contains information about emails exchanged between customer and sales representative
   l. Email-Wayfair: This class contains information about email subscriptions of Wayfair
2. Holdout data: This data includes all the features except the outcome variables for a different set of customers. We will use these features to predict the missing outcome variables.

III. Methodology

1. **Data Cleaning:** The dataset contains 10 categorical features and 170 numerical features. The first step is to check for missing values to give us a better understanding of how to work with this real-world data. Almost 70 columns had more than 65% missing values. Features with more than 80% missing values were dropped and the remaining were imputed using a median value.
2. **Handling Categorical Data:** The dataset contains a total of 10 categorical features and the general measures taken to transform those into numerical features were integer encoding and one-hot encoding.

   This involves two steps: Integer Encoding and One-hot Encoding. As a first step each unique category value is assigned an integer value and then the integer encoded variable is removed and a new binary variable is added for each unique integer value. We use one-hot encoding for all the other categorical variables. Some of the features affected were number of employees on the client side, number of purchases made by the client, cost of the purchases made by the client.

3. **Normalizing and Standardizing Data:** Next step is to basically shrinks the range of the data such that the range is fixed between 0 and 1(or -1 if there are negative values) to make the training less sensitive to the scale of features.
4. **Handling Class Imbalance:** Almost 89 per cent of the customers do not convert to new purchases in the next 30 days. There are several techniques explored in the case-study that can be used in dealing with class imbalance but none of them outperformed. In our case, oversampling the minority class would be a good choice because we don't have a ton of data to work with. Another method that we experimented with was undersampling the majority class. Undersampling can be defined as removing some observations of the majority class. But the drawback here could be that we might end up removing information that may be valuable or relevant. This could lead to underfitting and poor generalization to the test set, which was the case on this data set and so upsampling was a better idea. The figure below shows the distribution of the class convert_30 in our dataset.



5. **Feature-Selection Techniques:** This case study requires a lot of exploration with different feature selection techniques to extract the best performing feature subset. There are an overwhelming number of features and there is a need to make the model simpler for the interpretability purposes, to reason out and tackle the business problem in such a way that there is an explanation and logical reasoning behind the important features.
    a. **Filter Method:** In this method, we basically use Pearson's Correlation Coefficient to remove the correlated features. All the relevant features which were correlated to the target (more than 0.2) were extracted. Correlation between the features was checked and the highly correlated features were dropped to remove redundancy. This method outputs feature sets containing: 1) Days since last order 2) Number of online visits in the past 1 day 3) Number of online visits in the past 1-3 days 4) Number of online visits in the past 30-7 days 5) Number of online visits in the past 60-30 days 5) Number of ATCs (Add To Cart) in the past 7-3 days 6) Number of visits in the past year 7) Number of seconds on site in the past 7-3 days 8) Number of ATCc in the past 30-7 days 9) Days since last visit 10) Number of search terms in the past 30-7 days 11) rollup - unmanaged.
    b. **Wrapper Method:** The Recursive Feature Elimination (RFE) method works by recursively removing attributes and building a model on those attributes that remain. It uses accuracy metric to rank the feature according to their importance. The RFE method takes the model to be used and the number of required features as input. It then gives the ranking of all the variables, 1 being the most important. It also gives its support, True being relevant feature and False being irrelevant feature. This technique outputs the following features: 1) Percentage of dirty orders in the last thirty days 2) Conversion rate of quotes in the past year 3) Number of online visits in the past 1 day 4) Number of online visits in the past 1-3 days 5) Number of online visits in the past 3-7 6) Number of online visits in the past 7-30 days 7) Number of online visits in the past 30-60 days 8) Number of online visits in the past 60days to 1 year 9) Number of ATCs (Add To Cart) in the past 1 day 10) Number of ATCs (Add To Cart) in the past 1-3 days 11) Number of idea board adds in the past 1 day (Idea Board is a 'favorites' list) 12) Number of idea board adds in the past 3-7 days 13) Number of tasks (first introduction) in the past 30-60 days 14) Number of emails (between customer and sales rep) in the past 1 day 15) Percent of emails (from Wayfair) from the past 1 day that were opened16) Percent of emails (from Wayfair) from the past 3-7 days that were opened 17) Percent of emails (from Wayfair) from the past 60 days to 1 year that were opened 18) roll_up_Unmanaged 19) currentstatus_Enrolled 20) customersource_Internal Customer Scrape, 21) customersource_Social - Paid all of which are further discussed in the future section. Although there was some degree of similarity between the feature sets computed by the 2 techniques, we did discover some important feature sets from this method. We

fit a logistic regression model over the selected feature sets. The idea is to compare the models with a variety of different feature subsets. Notice here that we are not yet diving into the modeling part or using logistic regression for the purpose of prediction. We then see that the accuracy of the model when using all features in the dataset is 77% and when using 12 selected features is 75%. This does not tell us a lot about the subset. Although, it does not help in boosting the baseline performance, it is intuitive that most of the features in our dataset are not relevant. To further boost the performance, we explore some more feature selection techniques.

c. **Chi-squared:** Third method was based on computing chi-squared statistics between each non-negative feature and class. This score can be used to select the top n features features with the highest values for the test chi-squared statistic from X. Dataset contained a lot of negative values, but it was already processed to create a normalized dataset to perform Chi-squared tests. This particular method yielded the best performing set which was eventually revealed when different modelling techniques were applied. More so, later on it became a little intuitive to understand why this set performs better than the rest. It basically outputs 30 variables after which it starts to underperform. It combines all of the features from the previous 2 methods and as the number of features is more than the previous 2 methods - it turns out slightly better than the rest.
d. **Embedded Method:** Lasso Regularization- Lasso penalizes all of the features if it is not relevant. it penalizes the coefficient and makes it 0. Hence the features with a coefficient of 0 are removed and the rest are taken.

IV. INTERPRETABILITY OF FEATURE SUBSET SELECTION

Various feature selection techniques as described in the previous section were employed for identifying actionable features to predict and ideally boost conversion. While there is a lot of overlap in the relevant feature subsets, this section essentially discusses all of the relevant features from all of the methods and how they can be inferred in the business context to boost conversion.

First technique used was the ***filter method*** based on using Pearson's correlation coefficient. As the name suggests, we filter out the irrelevant features, select the relevant ones and then build the model. This model outputs many features around 'Number of Visits' like **Number of online visits in the past 1 day** , **Number of online visits in the past 1-3 days** , **Number of online visits in the past 30-7 days** ,**Number of online visits in the past 60-30 days.** While the Number of online visits metric does look a little repetitive, it does bring in different perspectives: It brings in the powerful idea of user segmentation - New users v/s Old users. This particular metric can be useful in determining and designing new strategies around new and old users. New visitors interact much differently than returning visitors. In most cases the new visitors try to get familiar with the brand, and figure out if the company is a credible or appropriate fit for their needs. In the context of Wayfair, this metric can definitely prove a lot more useful. When there is a large number of new visitors, one of the possible strategies could be to add them to their marketing list, send them promotional emails or give them a generous discount on their first purchase. In case of a large number of old visitors. When there are a large number of return visitors, it might be a good idea to tailor a personalized experience for them which is more consistent with their needs. Personalizing efforts to remind them of their previous purchases, product views, or interactions in general can help in boosting sales. Thus, this metric definitely does help in predicting purchase and if used wisely can help in channelizing efforts to convert the non-buying ones into the buying ones. The concept of a returning customer being easier to convert than a new customer can be validated by this set of relevant features as given out by the filter method.

Other important feature spit out by the model is: **Number of search terms in the past 30-7 days**. This again could logically be very important metric since it could essentially lay emphasis on site search: what the customers typically want and what pages aren't meeting their needs. If the number of search terms are low for a number of different products~ it probably means that search terms could be more specific and most likely coming from existing or returning customers who came back, knowing what they want. If the number of search terms are general and appear larger in number, then it may mean customers can't find what they are looking for. Another important perspective from this metric could be to test the Search Results Page. Are the customers able to find their items of interest on Page 1? It becomes very important to show the most relevant results *first.* Showing related products or recommended products to the returning customers can definitely pass the personalization test.For eg. showing Top picks. This is a very simple personalization tactic and can boost conversions. Furthermore, features like **Days since last order** and **Days since last visit.** These metrics could specifically be important in the context when return customers Customers visit the website, come back more times, depending on the price point and also on why the brand exists, and then maybe close the deal. These metrics also enhance the understanding of "how long" it takes for customers to buy from the website and is that behavior different across different segments of the website customers. If exploited the information wisely, it can definitely help optimize marketing campaigns, promotions, other efforts to boost conversion. Next, filter method places importance on **Number of Add To Carts in the past 30-7 Days -** While it is an intuitive one, there are many factors around this feature that can lead upto the actual conversion - Is there a guest check out option to enable faster processing (even if it's a bulk order)? Is there any shipping estimate provided when the product is added to the cart? Now that the customer has already made up his mind but is only a few clicks away - Are there any personalized efforts involved in reminding him to complete the purchase? Lastly, **Number of Seconds** on the website inarguably is another very important metric - How long is the website able to interest the user? Is the website search efficient in producing relevant results quickly? Again, is there any personalization to keep the user engaged? How is the product recommendation working. If the number is really low - we can definitely predict a higher bounce rate - The bounce rate is the rate at which new visitors visit the site and immediately click away without doing anything. Common problems may include poor website design, low usability, or higher load times that could potentially affect the conversion rates.

Second Feature selection technique was Recursive Feature Elimination and we get a number of features around various domains. While just like filter method, RFE Outputs **Number of ATCs** and **Number of visits** as important feature groups, it also places importance on customer satisfaction features like "**Percentage of orders in the past 30 days that were a 'dirty' (problematic) order**". E-satisfaction is definitely an important factor and lesser the number of problematic order, the greater the satisfaction. While this metric can be a direct metric by itself, there are a range of factors around this metric which can still help if the percentage of dirty orders is high. Customer obsession can be of the greatest value - How is the company tackling dirty order issues? how efficient and fast is the dirty-order and re-order processing? How does the company make sure it is not losing on its valuable customers? Is there any financial or product based incentive, especially because the case study concerns B2B customers and the orders associated with such customers are the bulk orders.

Next important feature is the "**Number of items in the favorite List**" since it can potentially lead to "**Number of items in the cart**". More so, another feature group identified via this method is "**Percentage of Emails Opened from Wayfair**" and "**Number of Emails between Customer and Sales Representative**". Email marketing is one of the most common forms of product marketing - often times A top priority for email marketing is to increase the subscriber engagement, which in turn has the power to increase sales, average revenue per customer and in general keep the customers informed about the new launches, on-going promotions and recommendations around their past purchases and in some way anticipate their future needs.

Chi-Squared test computes a set of features, most of which are a combination of features selected by the first 2 methods. This feature set performs better than the rest since it captures all of the important features from different feature subsets like "**Days since Last order**", "**Percentage of Dirty orders**", **Number of visits in the last few days**, "**Number of seconds on the site**", "**Number of Logged in sessions**", "**Number of ATCs**", "**Days since Last Visit**", "**Number of search Terms**" and "**Percentage of Emails opened**".

V. Modelling approaches and strategies

This business case involved a bit of experimentation with the model pipeline. Two approaches were tried and tested to meet the project goals.

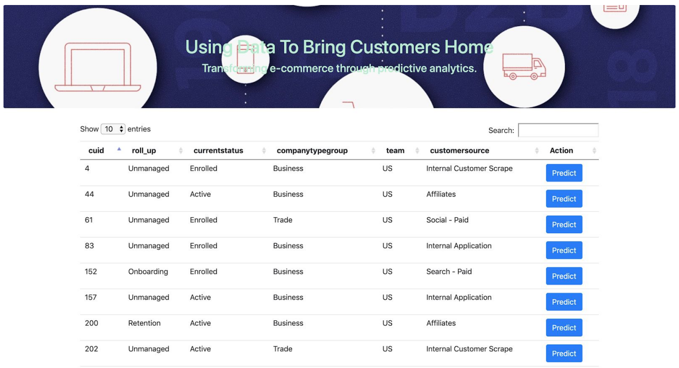
The first approach treats the classification and regression models independently. In this case, the two models independently predict whether the customer will convert or not and how much revenue can be generated from customers who do convert (do business with Wayfair) in the next 30 days. Both the models worked great independently with a balanced accuracy of 70% for the classification model and a good RMSE of $365 for the regression model. However, in spite of reasonable scores, the real problem incurred when the classification models predicts a customer will not make a purchase in the next 30 days, but the regression model predicts a purchase of some $$$ value. This discrepancy led to the evolution of another modelling approach. Thus, both the models did not have a way of communicating their results.

The second modelling approach treats this business problem as a whole. In this problem, we feed the values obtained from the classification model to the regression model. Thus, forming a pipeline between both models. We adhere to only those customers that actually make a purchase and predict the revenue for them. For the customers who do not convert we set a hard $0 value for them and do not predict a revenue using our regression model for them. Revenue_30 and convert_30 are highly correlated variables and the same feature set (top 30 features) are used for classification and regression problems.

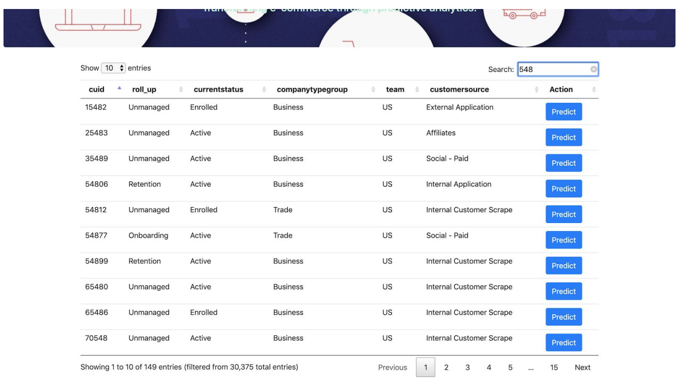
Classifier will predict whether a customer will make a purchase or not and the customers that actually make a purchase will serve as input to the regressor model where their revenue will be predicted. Note here that the classification and regression models, both are trained on complete train data sets. No bias has been introduced in any of the models by training only on customers that converted. To avoid the previously mentioned discrepancy, regressor model only predicts on those customers that actually make a purchase and outputs a $0 revenue for the customers that do not make a purchase.

VI. Web App

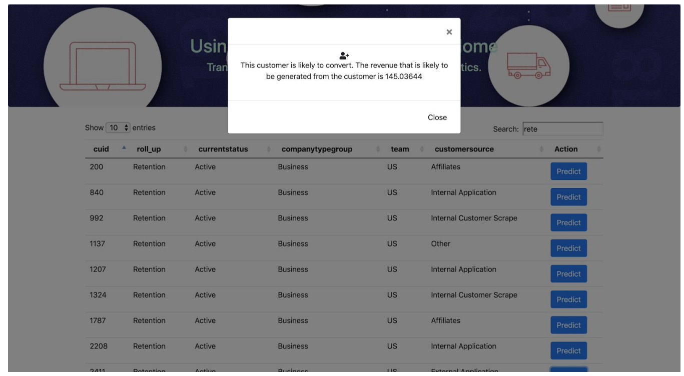
The final web application enables one to see the predictions made by the model. It tells you if a particular customer converts and if he does how much revenue can be generated from the particular customer. If a customer does not convert alongwith the message about the conversion it also shows key points which can be used to create targeted business strategies for this customer that can be used to convert these customers to make business. When one visits the web app, it is populated with the holdout dataset from Wayfair which did not contain information about these target variables. The figure below shows the landing page of the web app.

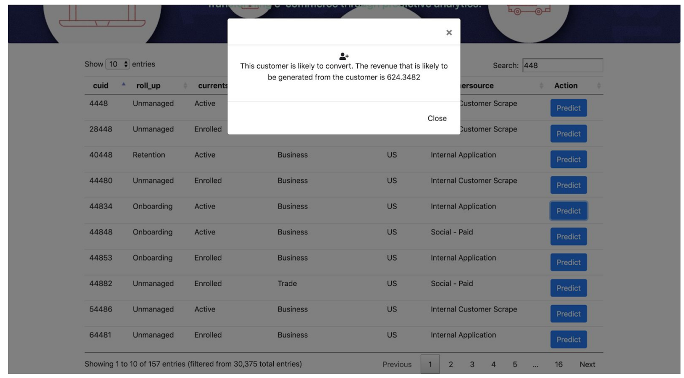


We use this dataset to make predictions with the help of our trained models. One can also search for a particular customer using the search bar on the home page. The figure below shows the search functionality in the web app.

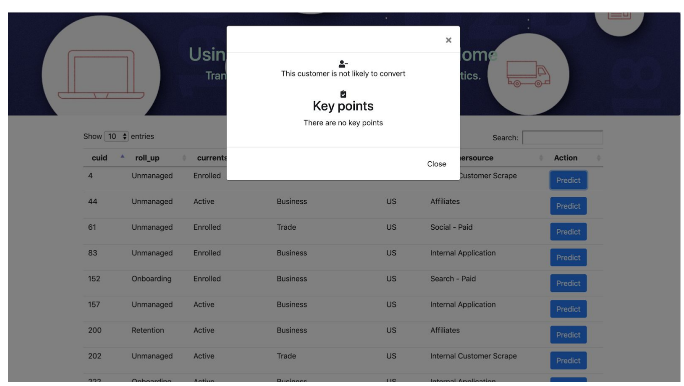


Once, a user clicks the predict button next to each customer, the platform begins recommending. The model will then predict if the particular user will buy something from Wayfair in the next 30 days. If he does, it also spits out a value of how much revenue can be generated from this customer. The two figures below shows a snippet of a few customers that did convert and the revenue that can be generated from these customers.





If a customer does not convert, alongwith the above information it also searches the feature values for this customer and checks if these values lie below a said threshold for that feature. The said threshold in our case are the z-scores we've obtained from the training dataset. If it's true then the name of this feature is shown as key points in the message box which might help creating targeted business strategies. The figure below shows a snippet of a customer that did not convert.



Note: The web app could not be hosted publicly as the dataset we used was a private dataset which was the property of Wayfair. In future if we do get permission from Wayfair, we might host the web app publicly. For now this it can easily be fitted on any other similar dataset and the web app can be recreated using the code on our Github repository.

VII. Results and Evaluation

The results that we obtained by modeling the different feature sets is given below. Inference from different feature sets has been discussed in the previous section.
For Logistic Regression we get:

Feature set obtained from	Accuracy	Recall
All 181 features	77	0.28
Filter method	75	0.27
Wrapper method	75	0.30
Chi-square tests	78	0.30
Embedded method	73	0.29

For Decision trees we get:

Feature set obtained from	Accuracy	Recall
All 181 features	64.20	0.33
Filter method	66.26	0.33
Wrapper method	63.18	0.31
Chi-square tests	76.53	0.46
Embedded method	65.49	0.33

For Random Forests we get:

Feature set obtained from	Accuracy	Recall
All 181 features	0.89	0.66
Filter method	0.88	0.65
Wrapper method	0.87	0.67
Chi-square tests	0.89	0.67
Embedded method	0.88	0.64

Furthermore, after the most predictive feature set was determined, a bunch of classification and regression models were trained, hypertuned and evaluated. Results are shown below.
In both cases, the extreme gradient boosting classification and regression models gave the best results.

Classifier	Accuracy	F1 score	Recall	PR	TNR	Balanced Accuracy
Logistic Regression	75.31	0.79	0.75	0.87	0.24	70.01
Decision Tree Classifier	76.54	0.8	0.76	0.87	0.26	70.34
Random Forest Classifier	89.05	0.86	0.89	0.85	0.47	0.55
Gradient Boosting Classifier	88.58	0.86	0.88	0.85	0.41	0.56
NBC	86.42	0.86	0.87	0.87	0.37	65.01
XGB	**82.00**	**0.82**	**0.79**	**0.87**	**0.28**	**70.27**

Regressor	RMSE
Linear Regression	1103
Random Forest Regressor	909
XGBRegressor	**507**

The results did not explain how the models were really performing. RMSE was not sufficient in understanding how the model is really performing. It did not give us the accurate idea of whether the right customers are being chased or not. The model at this point is not able to capture the revenue lost by not predicting the revenue for the customers who actually made a purchase. More so the model was not able to quantify the extra efforts that wayfair teams were spending on customers that are actually not making a purchase. To address these problems, there was a need to create custom business metrics for this use-case. The business metrics that were created are:

1) Business Loss: It is the total revenue that the model failed to capture because classifier input to the final regressor model was 0. Hence revenue predicted for those customers is 0.
2) Extra Efforts: The total revenue that is falsely predicted for the customers who did not convert.

The final model was then evaluated on test data set for all of the above mentioned metrics which yields the following results.

Balanced_Accuracy	0.73
Revenue_RMSE	$507
Extra_Efforts	23%
Business_Loss	24%

VIII. Statement of Contributions

The project is a work of joint efforts by both of the partners. While we divided the implementation part, all of the portions were jointly discussed,evaluated and inferred by both the team members. We collaborated online for code reviews, presentations and reports.

- Devanshi Gariba- Class Imbalance, Feature Selection, Modelling, Web App, Presentation, Report
- Alefiya Naseem- Data Cleaning, Wrangling, Class Imbalance, Modelling, Web App, Presentation, Report

References

[1] https://app.scholarjet.com/challenges/wayfairdata

[2] https://www.shopify.com/enterprise/44337411-how-to-increase-your-ecommerce-conversion-rates-using-these-3-analytics-reports

[3] https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18

[4] Chitsaz, Taheri, Katebi, Jahromi "An Improved Fuzzy Feature Clustering and Selection based on Chi-Squared-Test", Proceedings of the International MultiConference of Engineers and Computer Scientists 2009, Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong.

[5] Isabelle Guyon and Andr´e Elisseeff, 'An introduction to variable and feature selection', Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.

[6] A. Amin *et al.*, "Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study," in *IEEE Access*, vol. 4, pp. 7940-7957, 2016.

[7] https://nlp.stanford.edu/IR-book/html/htmledition/feature-selectionchi2-feature-selection-1.html

[8] Chawla, N. V. et al. "SMOTE: Synthetic Minority Over-Sampling Technique." Journal of Artificial Intelligence Research 16 (2002): 321–357. Crossref. Web.

Appendix

Link to Github code:

https://github.com/alefiya-naseem/CustomerRetention-Revenue

Model Pipeline Diagram