Executive Summary

Every time when a transaction occurs, there is a potential risk that the user is conducting a fraudulent activity, which negatively affects the operation of business. Hippo Inc., an E-commerce giant based in Canada, has recently experienced an increasing number of fraudulent activities, causing tremendous costs in chargebacks and loss of products to reduce the cost caused by fraudulent activities, Hippo Inc. has reached out to Thirty Thousand Consulting Inc. for further analysis. Hippo Inc. provided data from a similar anonymous company to Thirty Thousand Consulting Inc. The key business objective is to predict the probability that the first transaction of a new user is fraudulent, identify potential fraudulent activities and terminate these transactions to prevent fraud losses.

Thirty Thousand Consulting Inc. looks into the company's transaction data, performs data cleaning, such as outliers analysis and feature engineering, conducts exploratory data analysis, and builds logistic regression models. As a result, we find the number of times that a device is shared and the number of times that an IP is shared are the most important features to predict whether the transaction is fraudulent. Therefore, the recommendation is to limit the number of users to use the same device and the number of users to use the same IP. Besides, if the fraudulent probability is less than the safe ratio, the transaction will process successfully. If the fraudulent probability is higher than the safe ratio, but lower than the alert ratio, the transaction will require further verification, such as text verification. If the fraudulent probability is higher than the alert ratio, the transaction will be put on hold immediately. In this way, Hippo Inc. will experience fewer fraudulent activities.

Introduction

Over the last decade, we have seen a shift in consumer shopping behaviour from offline to online which led to the rise of e-commerce platforms. The increase in popularity of such transactional platforms created numerous opportunities for businesses as well as fraudsters. The risk of money laundering and fraudulent transactions with stolen credit cards are problems that all e-commerce platforms face and Hippo Inc. is no exception. Hippo Inc. as a Canadian e-commerce giant connects businesses and customers from all around the world and provides products ranging from furniture, electronic devices to designer clothing. Unfortunately, the company has recently experienced an increase in fraudulent activities as reported by its users. Fraudulent transactions eat away profit with chargeback fees, shipping expenses and lost merchandise, as well as resources used to handle them. Moreover, will negatively impact the brand image of Hippo Inc. perceived by loyal and potential users.

Proposal

Through preliminary analysis of the situation, Thirty Thousand Consulting Inc. found that conventional approaches to detecting fraud are outdated compared to increasingly sophisticated tools adopted by fraudsters, therefore is proposing a new comprehensive fraud management plan. As part of the plan, Hippo Inc. is recommended to adopt advanced algorithms that more efficiently and accurately detect fraudulent transactions in real time. To start with, it is suspected that most fraudulent activities were conducted by new users during their first transaction. Historical transactional data will be collected and cleaned, followed by exploratory data analysis. A baseline logistic regression model will then be built which predicts the probability that the first transaction conducted by a new user is fraudulent. The logistic regression model will be updated with an optimized probability threshold. An appropriate threshold value is important to both model performance and business implications because it should seek to balance fraud prevention efforts and user experience. In conclusion, it is critical for Hippo Inc. to adopt the fraud management plan proposed as it will help prevent loss from fraudulent activities and create a positive customer experience.

Overview of Data

The dataset was provided by an anonymous company, which includes 151,112 observations without missing values. There are 12 attributes in total after we merged the datasets properly. These attributes include 11 predictor variables and 1 response variable (fraud=yes/no). Please refer to the data dictionary (figure 1 in the appendix) for detailed information.

Methodology

Data pre-processing

1. Missing values and Outliers:
   We conducted descriptive statistical analysis on our raw data. Based on our analysis, we cautiously dealt with outliers and missing value.
2. Continuous and categorical variables:
   We have 7 continuous variables and 3 categorical variables. Categorical data (*Source, Browser, Sex*) were transformed to dummy variables using one hot encoding. From previous analysis, we found that the time variable *week of year* was related to fraudulent activities.
3. Oversampling and data partition

The dataset was split into training and testing sets and the training set accounted for 80% of total observations. Since the fraudulent transaction rate was only 9%, the training set was balanced by oversampling the fraudulent transaction observations.

Feature Engineering

During the process of exploratory analysis, we created some new attributes that were essential in tackling the topic at hand (Please refer to the submitted Proposal). Below are the new attributes created and kept:

1. 'signup_weekofyear': week of the year at time of sign up
2. 'purchase_weekofyear': week of the year at time of purchase
3. 'signTOpurchase': time difference between sign up and purchase
4. 'device_shared': number of times a device is shared
5. 'ip_shared': number of times a IP is shared
6. 'country_shared': number of times a user visited the website from a particular country

Exploratory Data Analysis

We found patterns and trends in the data which became the basis for further examination during the second phase of engagement, including but not limited to, in-depth data analysis, model tuning and evaluation. We divided all variables into three distinct types: numerical, categorical and time variables. For numerical variables, we compared the distribution of data for fraud and non-fraud records using histograms and density curves. For categorical and time variables, we analyzed them using count plots. Please refer to the submitted Proposal of the key findings obtained from the dataset.

Model Building and Evaluation

We chose logistic regression due to the fact that logistic regression shows superior when the dependent variable is dichotomous. Compared with black box models, logistic regression can give explicit coefficients, which helps interpret the model meanings and reveal how these variables drive fraudulent transactions.

- ❏ Baseline model:
  Model building process started with using ordinary logistic regression
- ❏ Regularization:
  We built ridge and lasso logistic regression to improve on the baseline model to avoid overfitting and enhance the accuracy. Ridge logistic regression turned out to fail in outperforming the lasso regression.

- ❑ Optimize the threshold probability:
  To optimize the threshold for lasso logistic regression, we want a sensitivity equal to specificity. Cross-Validation Method is used to get the best threshold. This threshold can be applied to the model when our client wants to be more sensitive when detecting the fraudulent transactions.

We evaluated our model based on multiple metrics, including accuracy, recall, precision and F1 score. Our final model here is lasso logistic regression with the accuracy of 92% and the recall of 72%. Depending on the risk preference, our clients can also apply different cut-off probabilities when identifying the fraudulent transactions. Based on statistical criteria, we would recommend the threshold to be 0.15 (when the predicted probability is larger than 0.15, it will be identified as fraud) to meet more risk averse needs.

Results

The final model was tested on the test set yielding 92% accuracy. For the result of lasso logistic regression, there are 18 coefficients corresponding to 17 independent variables and 1 intercept. According to the feature importance plot (figure2), we found that “age”, “country shared”, “sign to purchase”, and “signup_weekofyear” have almost no relationship with the log-odds of the dependent variable “fraud”. There are 3 positive coefficients and 10 negative coefficients. Positive coefficients imply that larger variables or these categories increase likelihood of fraud, while negative ones suggest the opposite. For detailed results of our final model, please refer to Figure 5 and Figure 6 in appendix.

Model Insights

From our model output, it is clear that both device shared and IP shared are the most important factors in determining whether the transaction is fraudulent. It is no surprise. The more times that a device is shared, it indicates that multiple accounts used the same device to purchase items on Hippo’s website. Normally, a regular person will not have multiple accounts. Hence, malicious users are using multiple accounts to scam. New IP address has a higher chance of making a fraudulent transaction. As for IP shared, the more unique an IP address is, the higher probability that the purchase will be fraudulent. This comes down to the fact that malicious users will use multiple IP addresses to avoid authority tracking onto them. It seems that the type of browser that users used to visit Hippo’s website is a strong indicator for fraudulent activities. IE browser users tend to have a higher fraudulent probability. As for source of traffic, SEO and Advertisement access seems to be responsible for most of the fraudulent cases. Female users seem to conduct more fraudulent acts than male. However, malicious users might misrepresent their identities.

Besides the predictive model we provided for our client, we also designed an algorithm that outlined the trade-off between sensitivity and specificity such that our client can choose the best trade off that matches their risk preferences.

Recommendation and Conclusion

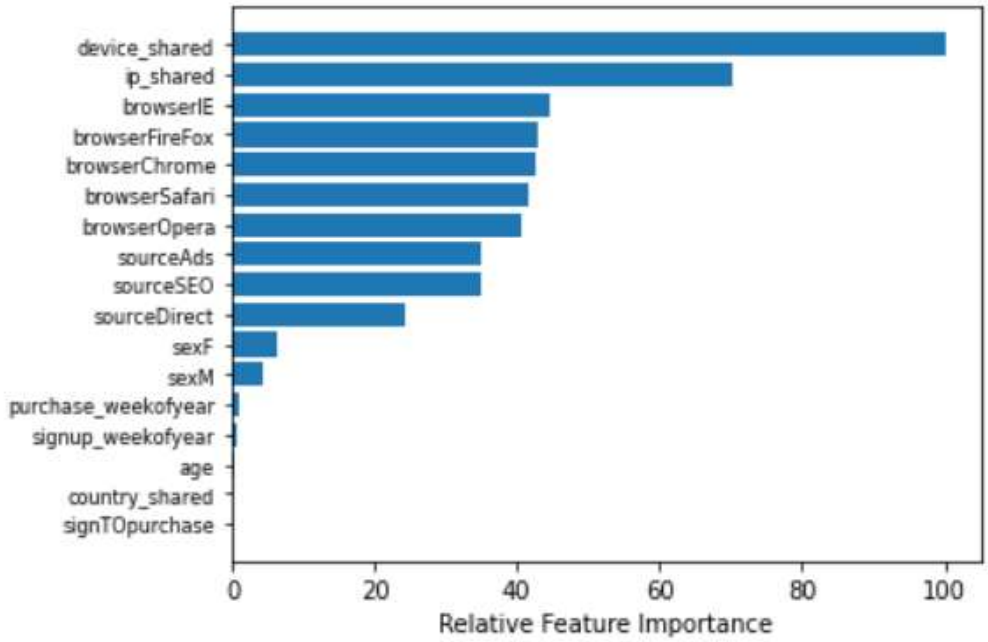
In order to predict the probability that the first transaction of a new user is fraudulent and identify potential fraudulent activities, we conduct logistic regression models to predict the fraudulent activities and make decisions based on those ratios. Consequently, we recommend the following steps for Hippo Inc. to implement to prevent future fraudulent transactions:

- ❏ Limiting the number of users per device/IP
    - ❏ As the results show, the number of times that a device is shared and the number of times that an IP is shared are the most important features to predict the fraudulent probability.
- ❏ Banning IP/Device ID associated with fraudulent activities
    - ❏ IP address that conducted a fraudulent activity before will have that same IP address permanently banned from accessing the website again. Same can be applied for device ID.
- ❏ Processing transactions based on probability
    - ❏ If the fraudulent probability is less than the safe ratio, the transaction will process successfully. If the fraudulent probability is higher than the safe ratio, but lower than the alert ratio, the transaction will require further verification, such as text verification. If the fraudulent probability is higher than the alert ratio, the transaction will be put on hold immediately.

Appendix

data fields	definition
signup_time	the time when the user created her account (GMT time)
purchase_time	the time when the user bought the item (GMT time)
purchase_value	the cost of the item purchased (USD)
device_id	the device id. You can assume that it is unique by the device. I.e., transactions with the same device ID means that the same physical device was used to buy
source	user marketing channel: ads, SEO, Direct (i.e. came to the site by directly typing the site address on the browser).
browser	the browser used by the user.
sex	user sex: Male/Female
age	user age
ip_address	user numeric ip address
country	country where user is logged in
class	this is what we are trying to predict: whether the activity was fraudulent (1) or not (0).

(Figure1: Data Dictionary)



(Figure2: Feature Importance Plot)

	precision	recall	f1-score	support
0	0.97	0.94	0.95	34286
1	0.54	0.72	0.62	3492
accuracy			0.92	37778
macro avg	0.76	0.83	0.79	37778
weighted avg	0.93	0.92	0.92	37778

(Figure 3: Evaluation metrics for lasso logistic regression with default threshold)

	precision	recall	f1-score	support
0	0.97	0.76	0.85	34286
1	0.25	0.78	0.38	3492
accuracy			0.76	37778
macro avg	0.61	0.77	0.62	37778
weighted avg	0.90	0.76	0.81	37778

(Figure 4: Evaluation metrics for lasso logistic regression with threshold = 0.15)

The function of Lasso logistic regression:

$$Log(p/(1-p)) = b0 + b1*x1 + b2*x2 + \ldots\ldots + b16*x16 + b17*x17.$$

(Figure 5: Equation of Final Model)

Variables(x_i)	**coefficients(b_i)**	**Variables(xi)2**	**Coefficients(bi)2**
signup_weekofyear	1.35E-02	sourceSEO	-6.91E-01
purchase_weekofyear	-1.61E-02	browserChrome	-8.43E-01
age	-5.17E-03	browserFireFox	-8.49E-01
signTOpurchase	3.97E-09	browserIE	-8.80E-01
device_shared	1.97E+00	browserOpera	-8.04E-01
ip_shared	-1.38E+00	browserSafari	-8.18E-01
country_shared	-5.25E-07	sexF	-1.22E-01
sourceAds	-6.92E-01	sexM	-8.32E-02
sourceDirect	-4.81E-01		

(Figure 6: Coefficients of Final Model)