Customer flight satisfaction prediction

Foundations of Data Science, 2020/2021 project

Chicca Lorenzo 1708956, Choma Patryk 1617474, D'Evangelista Nicolò 1698229, Hysaj Rigels 1706263

Abstract

Our project aims to predict if a flight passenger is going to be satisfied of the service offered by the airline or not. Since the prediction is a binary value, we decided to make use of Logistic Regression to accomplish this goal. The parameters we find after the training manage to predict correctly with an ~80% accuracy rate.

Introduction

In order to improve the satisfaction of customers during flights, we thought about implementing an algorithm able to predict whether a passenger will be satisfied from a flight or not.

Related Work

We took inspiration from a *response*[1] to the dataset we used from *kaggle,* which explores the results obtained by using various models on the dataset. We imitated his choice of feature, eliminating the ones least correlated with "satisfaction".

Method

We chose to use Logistic Regression because of the binary nature of our target and because of our familiarity with the method.

In particular we obtain predictions using Gradient Ascent, Newton's method and Gaussian Discriminant Analysis, and we compare them using the area under the ROC curve to determine which method makes the best metric.

Dataset and Benchmark

The dataset used is taken from *kaggle*[2], which has some very desirable qualities: it's pretty big, containing more than 120,000 rows, with only a few missing values, and mostly quantitative data, making it easy to work with.

We pre-process the data by eliminating the features least correlated with "satisfaction", encoding the qualitative features (using One-Hot Encoding) and as for the rows with missing values, since they are less than ~400, therefore we decided to delete this miniscule percentage of the whole dataset. Since the "age" feature is on a completely different scale than the others, the convergence is really slow and requires really small learning rate. Therefore, we rescale all the features into the [0, 1] range, resulting in very fast convergence.

Finally, we decided to split the dataset into 70% training, 30% testing, which is more than enough training data considering it's ~90,000 rows.

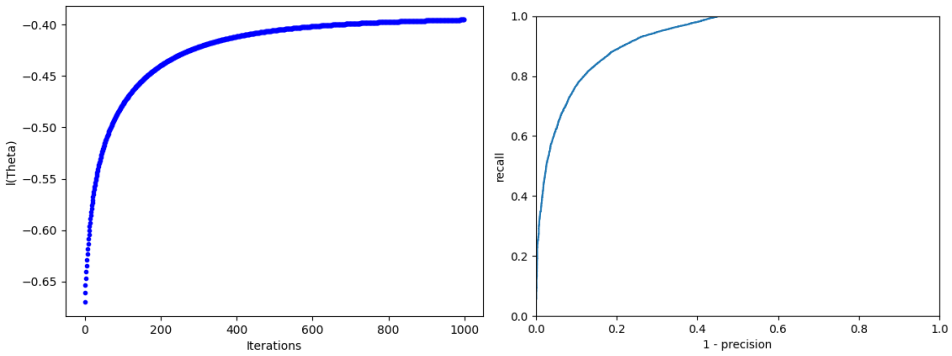
Experimental results

When using Gradient Ascent, we converge pretty fast, within 1000 iterations with a learning rate of 0.3. As we can see from the precision-recall curve, which covers most of the graph area, the resulting predictions make a good metric for deciding if a customer will be satisfied or not. The achieved accuracy with this method is 83%.

---

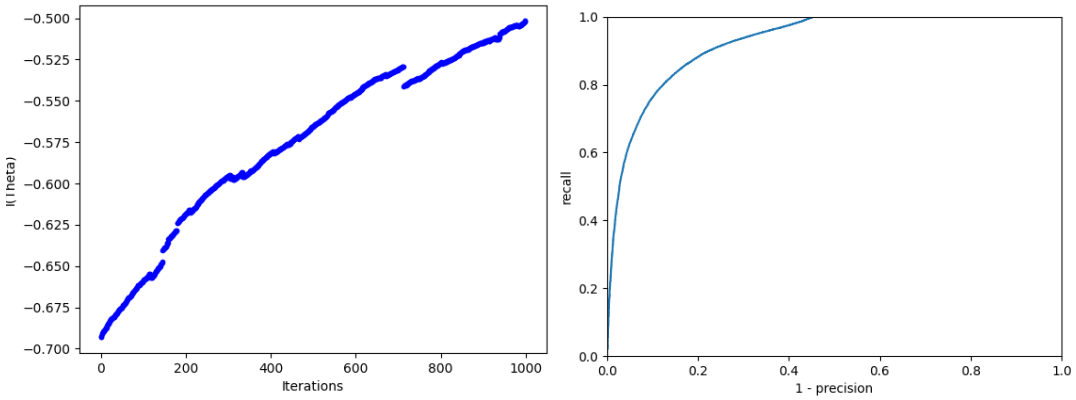
[1] https://www.kaggle.com/bimarshakhanal/airlines-customer-satisfaction

[2] https://www.kaggle.com/sjleshrac/airlines-customer-satisfaction

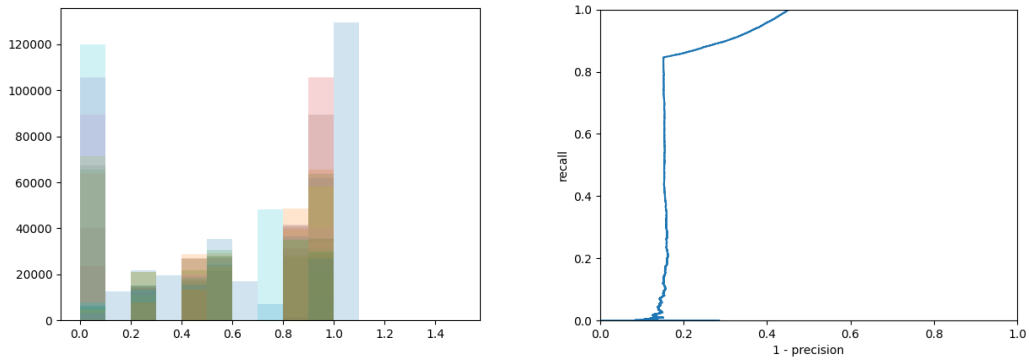


When using Newton's method, we also use 1000 iterations, but we only use 1% of the dataset (~1200 random rows), for various reasons: most importantly, computation time increases dramatically because of the huge matrices involved, and more often than not we encounter singular matrices, of which we cannot take the inverse for the method.

We managed to achieve an accuracy of 80%, but this was mostly due to a lucky initialization, partially caused by the small percentage of the dataset used. Newton's method may not be the best for this case, as Newton's method tends to converge to *saddle points*, which are prevalent in high-dimensional spaces[3]. As soon as we change the random seed for the train/test split we get very different results.



When using GDA we achieve the best accuracy, with a rate of 83.3% correct predictions when training with 70% of the dataset, but the area under the roc curve is the worst out of the three (especially when considering our specific case for Newton's method). Despite this, the accuracy obtained with this method tends to be very high even when using just 1% of the dataset (almost 83%) and even when using as few as 0.1% of the dataset, it manages to correctly predict 76% of the testing data. We plotted all variables' distributions against each other, and found that not all of them follow a normal distribution. This could lower the performance of this method, but the predictions obtained are still fairly accurate.

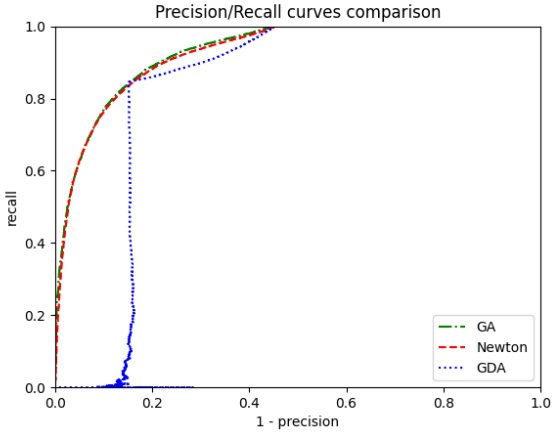


[3] Identifying and attacking the saddle point problem in high-dimensional non-convex optimization

Conclusions

In conclusion, we obtain the best results for accuracy when using GDA, closely followed by Gradient Ascent. As we can see from the RoC comparison graph, GDA seems to be a poor metric in comparison, but its top-left corner peaks above the other two. This should correspond to the point for which the threshold is equal to 0, which we use to calculate accuracy.

Therefore, Gradient Ascent seems to be the most reliable metric for all threshold values, and GDA slightly outperforms it when the threshold is exactly 0.



References

1 https://www.kaggle.com/bimarshakhanal/airlines-customer-satisfaction

2 https://www.kaggle.com/sjleshrac/airlines-customer-satisfaction

3 Identifying and attacking the saddle point problem in high-dimensional non-convex optimization