Hackathon 2023 - 1.2.3

Amit Benbenishti, Adi Shnaidman, Shoham Mazuz, Omer Ventura

Introduction

The churn prediction model aims to identify the most relevant features for predicting cancellations and flagging orders at risk. The objective is to create a feature set that is informative and minimal, maximizing predictive power while minimizing complexity.

Data Preparation and Feature Selection Methodology

In the process of developing our churn prediction model, we recognized the critical importance of selecting the most informative features for accurately predicting cancellations. To achieve this, we employed a robust feature selection methodology that involved utilizing ensemble and loadings from Principal Component Analysis (PCA). This combination of techniques allowed us to comprehensively evaluate the individual performance of each feature while capturing the underlying patterns in the data.
An ensemble, which evaluates each feature individually, provided us with a straightforward and interpretable way to assess the predictive power of each feature. On the other hand, by considering the loadings of the features, we gained insights into their contribution to the overall variance of the data. By leveraging both ensemble and PCA loadings, we were able to capture the distinct strengths of each approach, resulting in a comprehensive feature selection methodology.

Selected Features

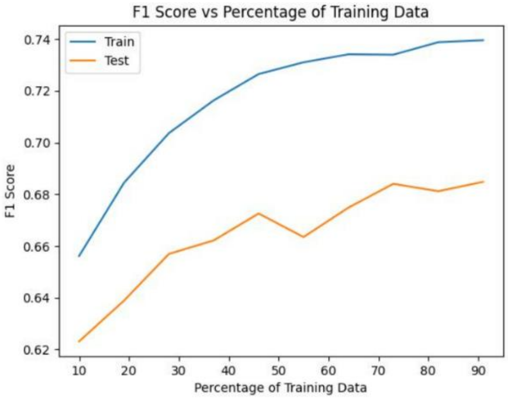
Through our feature selection methodology, we successfully identified the top four features with the highest predictive power for cancellations. These features, namely 'days_ahead', '>30', '<1', and '2-3', emerged as the most significant contributors to our churn prediction model. It is noteworthy that all four features were created during the preprocessing phase.

The feature 'days_ahead' represents the difference in days between the check-in date and the booking date.
The feature '>30' serves as an indicator that triggers a penalty payment if the reservation is canceled within a range of 30 days before the check-in date.
Similarly, the feature '<1' functions as an indicator that denotes the requirement of a penalty payment if the reservation is canceled within a day before the check-in date.
Lastly, the feature '2-3' serves as an indicator for penalty payment if the reservation is canceled within 2-3 days before the check-in date.

By incorporating these four features into our churn prediction model, we can effectively leverage the predictive power they offer. Their inclusion enhances the model's ability to flag orders at risk and make accurate predictions regarding cancellations. The significance of these features underscores their valuable contribution to the overall effectiveness of our churn prediction model.
By employing ensemble learning and PCA, we identified the four most significant features for predicting cancellations. To assess their importance, we conducted an evaluation using a dataset consisting solely of these four features. Remarkably, the F1 score macro obtained from this subset closely matched the F1 score achieved on the entire preprocessed test data. Specifically, the subset yielded an F1 score of approximately 0.695, while the overall F1 score stood at approximately 0.71. This compelling result indicates that these four features alone possess a remarkable ability to capture the relevant patterns and information necessary for making accurate predictions about cancellations.

Graphs analysis

This following graph plots the F1 score on the y-axis, which indicates the new model's (consisting of only the 4 significant features) performance, against the percentage of training data used on the x-axis ranging from 10% to 90%.



The graph shows two lines, one showing the performance of the 4 significant features over the training data and one showing it's performance over the test-data.

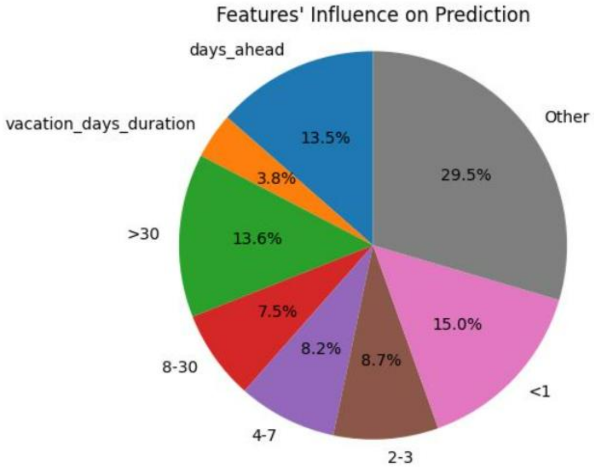
The graph provides a visual representation of how the model's performance improves and stabilizes with an increasing amount of training data. It is evident that as more training data is used, the model's performance becomes more consistent and reliable.

Additionally, the graph reveals that the performance of the new model on both the training set and the test set is remarkably similar. This suggests that the model generalizes well and is able to perform consistently on unseen data, which is a desirable characteristic for a machine learning model.

Furthermore, it is worth noting that the performance of the new model is impressively close to that of the previous model. This indicates that the new model has been able to achieve a comparable level of performance, potentially even surpassing it. This is a positive outcome and demonstrates the effectiveness of the new model in capturing and learning from the underlying patterns in the data.

This following pie chart represents the influence of different features on the prediction made by the model. The chart shows the relative importance of each feature in the model's prediction.

The following data was set by values taken from sklearn's model, representing each feature weight over prediction.



The features that have an influence above a certain threshold (0.03 times the total sum) are considered significant and included in the chart individually. Any feature with an influence below the threshold is grouped together under the "Other" label.

This graph shows that the features previously mentioned, are in the significant ones.

Conclusion

In conclusion, our feature selection methodology, which combined ensemble and PCA, proved to be highly effective in identifying the most significant features for our churn prediction model. The comprehensive evaluation of individual feature performance and underlying data patterns allowed us to confidently select the top features with the highest predictive power. The close alignment between the F1 scores of the selected feature subset and the overall preprocessed test data further reinforces the significance of these features in contributing to the overall predictive capability of our churn prediction model. By leveraging the strengths of decision stumps and PCA, we have equipped our model with the most informative variables, ensuring its efficacy in identifying and flagging orders at risk of cancellation.