NOVA IMS

Data-Driven Decision-Making

Fernando Bação **Farina Pontejos** **Ivo Bernardo**

CLASSIFICATION - HOTEL CANCELLATION PREDICTION

GROUP 10:

Eduardo Xavier, 20230355

Joana Oliveira, 20230384

Joana Franco, 20230365

João Chang, 20201620

Rúben Machado, 20230367

JUNE 2024

TABLE OF CONTENTS

1.EXECUTIVE SUMMARY

The Lorem Ipsum Hotel Group requested that we develop an updated approach to managing cancellations. As data-driven consultants, our mission is to develop a predictive model to address this issue. The objective of this initiative is to enhance revenue management by accurately forecasting cancellations and optimizing resource allocation. By employing data preprocessing, feature selection, and model evaluation techniques, we were able to construct and validate predictive models utilizing KNIME Analytics software.

Two machine learning models were trained and evaluated: a Decision Tree and a Random Forest.

In order to optimize the efficacy of the model, it is recommended that the financial impact of missed versus false cancellations be analyzed and that the model sensitivity be adjusted accordingly. Furthermore, the model should be updated on a regular basis with new features, in order to capture evolving trends and maintain prediction accuracy.

Adoption of this data-driven approach will enable the Lorem Ipsum Hotel Group to significantly enhance its revenue management strategies, ensuring efficient resource allocation and improved guest experiences. This transition will facilitate the development of a more responsive and sustainable business model, which will ultimately contribute to long-term success.

2.BUSINESS AND DATA UNDERSTANDING & DATA PREPROCESSING

Business and Data Understanding

The dataset comprises 72646 rows and 25 variables regarding hotel cancellations. The initial phase of this project involved an analysis of the data and the subsequent steps to be followed. This was achieved through the use of heatmaps and bar charts (Appendices 2 and 3), which were employed to analyze the descriptive statistics and structure of the dataset.

Firstly, the IsCanceled and IsRepeatedGuest variables were transformed into a string format as they are binary variables, with 0 corresponding to ‘no’ and 1 to ‘yes’.

Upon examination of the statistics view node (Appendix 1), it was verified that there are 11 string-type variables, 12 integers, 2 doubles, and 1 variable with a local date type.

There are no variables with unique values and no redundant variables. The Agent and Company columns displayed 'NULL' values, therefore we proceeded to use the "String Replacer" node to transform these values into missing values. Furthermore, the "String to Number" node was employed to correct the erroneously assigned data types in the respective columns.

Regarding the missing values, the only variables exhibiting any missing records were Children with 4 and Country with 24.

Given that the maximum number of babies displayed was 10 and the Average Daily Rate (ADR) had a maximum value of 5400 monetary units, we became aware of the potential for the presence of outliers.

Data Preprocessing

In the Data Preparation metanode, the variables BookingID, Company and Agent were excluded from the train and test data. The variables are all ID's, which is the primary reason for their removal. Additionally, Company was excluded due to the high number of null values (69 095).

In order to ensure the model's efficacy and to prevent overfitting, the partitioning node was employed on the training data, with a 70/30 split. As the objective is to apply the model to test data in order to obtain predictions, it is unnecessary to split the Hotel test data. For the missing values in both datasets, the mean was applied to the numeric values and the most frequent value was applied to the strings. Some variables exhibited outliers, such as the Babies variable, which had two observations with values of 9 and 10. Furthermore, the replacement strategy of closest permitted value was employed for both datasets in order to address these and other outliers.

The Hotel Data and Hotel Test datasets include variables with different ranges, which may have a negative impact on the model and predictions. To address this issue, Min-Max Normalization was applied, ensuring that all variables have a range of 0 to 1. This normalization process is crucial for improving the learning process and overall model accuracy, as it ensures that all variables contribute equally. Normalizing the Hotel Test data helps us to make sure that the training and testing conditions are identical, thereby facilitating the generation of accurate predictions.

Upon examination of the Correlation Matrix (Appendix 4), it was determined that the variable DistributionChannel should be removed, as its correlation with the variable MarketSegment exceeded the previously defined threshold of 0.7, reaching 0.7338.
Afterwards, we filtered out the columns Meal and RequiredParkingSpaces as they were deemed to be irrelevant to predict the target.

3.MODEL ENGINEERING AND EVALUATION

In an effort to get the optimal result in the prediction of the target variable on the test dataset, two different predictive models were trained on the train set: a Decision Tree and a Random Forest.

Decision Tree

We decided to implement a Decision Tree considering the flexibility of its implementation, its robustness to outliers, and its ability to solve complex problems due to its nonlinear approach.
The model's evaluation metric indicated an accuracy of 81.05% (Appendix 6), which signifies its performance in correctly predicting outcomes (Appendix 5A).

Random Forest

In addition, we decided to implement a Random Forest because of its versatility and robustness to overfitting.
The model's evaluation metric indicated an accuracy of 80.619% (Appendix 6), which signifies its performance in correctly predicting outcomes (Appendix 5B).

Binary Classification Inspector

The Binary Classification Inspector node was employed to assess the performance of the Decision Tree and Random Forest models based on different performance metrics (Appendix 7).
ROC, which stands for Receiver Operating Characteristic, is a graphical plot that illustrates the trade-off between True Positive Rate and False Positive Rate at various classification thresholds. AUC, which stands for Area Under the Curve, is a single metric that represents the overall performance of a binary classification model based on the area under its ROC curve. A higher AUC indicates that the model is more effective at predicting 0 classes as 0 and 1 classes as 1.

Another metric that is also used in Appendix 7 is the Accuracy, which only considers the number of correct predictions overall.

Consequently, we elected to utilize the ROC AUC evaluation metric, which offers a more robust indicator of model performance.

The ROC curves of the Random Forest (green) and Decision Tree (blue) models were compared. This revealed that along the y-axis (representing the sensitivity or True Positive Rate), the blue line exhibited a superior identification of the positive cases (85% against 67%). On the x-axis (representing False Positive Rate, the inverse of Specitivity), the green line misclassified negative cases on only a few occasions (0.95% against 9.60%).

The analysis of the AUC of both models revealed that the Decision Tree covers a wider area, indicating slightly better performance. Given that the Decision Tree model performs better on both metrics, it is recommended that the Hotel implement this model as a cancellation management approach.

4.DEPLOYMENT

In light of the confusion matrix presented in Appendix 5A, it can be observed that the True Positives (10664) represent the bookings that were correctly identified as not being canceled (predicted as 0). This is a favorable outcome for the hotel, as it demonstrates the effective utilization of its resources. The true negatives (7000) represent bookings that the system correctly identified as cancellations (predicted as 1). This is a favorable outcome for the hotel, as it allows them to adjust their staffing levels and resources, or potentially resell the room.

The false positives (2197) represent bookings that were not predicted as cancellations (predicted as 0) but were in fact canceled (predicted as 1). This may have adverse implications for the hotel. For example, the hotel may lose revenue due to the unfilled room. Furthermore, this can have an adverse effect on the guest experience, as those who have made the journey to the hotel in the expectation of occupying a room may be denied access, which could result in a negative perception of the hotel and the likelihood of a negative review. Furthermore, the hotel may be forced to allocate resources ineffectively, as staff and resources may be allocated to bookings that do not materialize.

The False Negatives (1933) represent the bookings that were predicted as cancellations (predicted as 1) but did not actually get canceled (predicted as 0). This approach, however, can also have drawbacks. For instance, it may result in the inefficient use of resources, as the hotel may allocate unnecessary resources for cancellations that do not materialize in order to mitigate the inconvenience of the situation. Additionally, financial costs are incurred due to the necessity of paying for the relocation of guests to alternative accommodations. Furthermore, this can impact the guest experience, as those who have received cancellation notices in error may experience confusion or frustration. Such an outcome may result in a negative experience for the guest, leading to a decline in future customer loyalty and potential bookings.

In light of the trade-off between missed cancellations (FP) and false cancellations (FN), it is recommended that the following measures be taken in order to improve the performance and revenue of the hotel.

Firstly, the hotel should undertake a cost-benefit analysis to determine the average revenue lost per missed cancellation (FP) and the cost incurred per false cancellation (FN) due to resource waste. This will assist the hotel in comprehending the financial implications of each type of error.

Subsequently, the model sensitivity should be adjusted based on the results of the cost analysis. If the cost of missed cancellations (FP) is significantly higher than the cost of false cancellations (FN), the hotel can adjust the model to be more sensitive.

This will likely result in an increase in the number of false negatives (FN), predicting more cancellations than actual, but a reduction in the number of false positives (FP), predicting more cancellations than actual. One may consider adjusting the threshold of the employed model or utilizing cost-sensitive learning algorithms as common techniques.

Furthermore, the hotel can utilize the model's predictions to prioritize follow-up actions for bookings with a high likelihood of cancellation (predicted as 1 with high probability). For example, the hotel may wish to consider contacting these guests with targeted promotions or incentives in order to encourage them to retain their reservations. Alternatively, the hotel could offer alternative dates or room options if necessary. For bookings with a low likelihood of cancellation (predicted as 0 with high probability), the optimal allocation of resources can be determined.

5.MONITORING AND MAINTENANCE

In order to maintain the model's quality, it is recommended that the hotel should keep it updated by periodically examining its principal features.

While the current data set likely includes factors such as meal, customer type and required parking spaces, the inclusion of additional features could enhance the accuracy of the predictions.

TotalSpecialRequests: The number and type of special requests made by a customer. For example, information regarding room preferences, early check-in requests, and dietary needs may indicate an elevated risk of cancellation if these requests cannot be readily accommodated.

CustomerReviews: The analysis of customer reviews can provide insights into customer satisfaction and potential cancellation triggers (sentiment analysis).

The analysis of reviews for negative sentiment or frequent mentions of specific concerns (e.g. cleanliness, noise) may serve to identify bookings with a higher cancellation risk.

NumberRoomsRequested: Bookings for multiple rooms, particularly those made by new customers, may be indicative of a heightened risk of cancellation, given the potential for plans to evolve nearer to the date.

GuestSatisfaction: A rating on a scale of 1 to 10 (from detractor to promoter) is given for the overall hotel stay and experience, according to the NPS scale.

In addition to the aforementioned models, other approaches may be considered, such as Naive Bayes, due to its simplicity and efficiency. This particular model is an effective approach as it is computationally inexpensive to train and requires relatively low memory compared to other algorithms.

6.CONCLUSION

During the course of our project, we encountered a number of challenges, including the presence of incomplete and inconsistent data, the need to determine which features were most meaningful in predicting our target, and the task of selecting the most appropriate model for accurately capturing cancellations.

Our analysis indicated that there was a substantial opportunity to enhance hotel revenue management through the implementation of a data-driven approach to cancellation prediction. The existing model provides insights that are valuable in their own right. However, its performance can be further enhanced through the application of AI, with a particular focus on improving the customer experience. This could be achieved by collecting an immediate survey at the point of check-out. Furthermore, the hotel can create user-friendly dashboards to assist decision-makers.

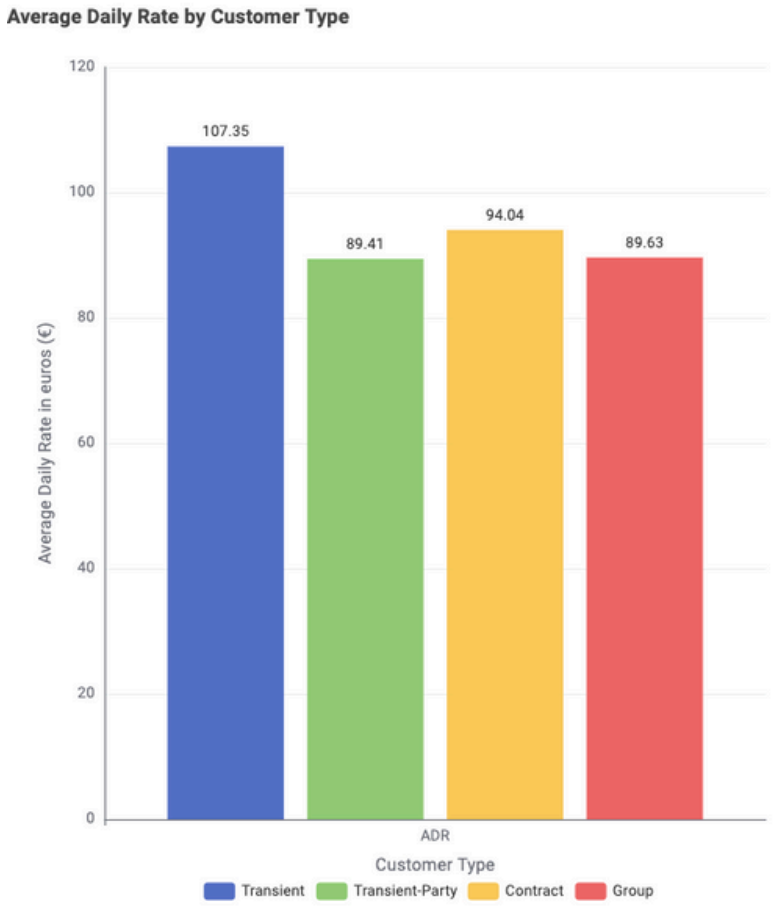
7.APPENDICES

Statistics

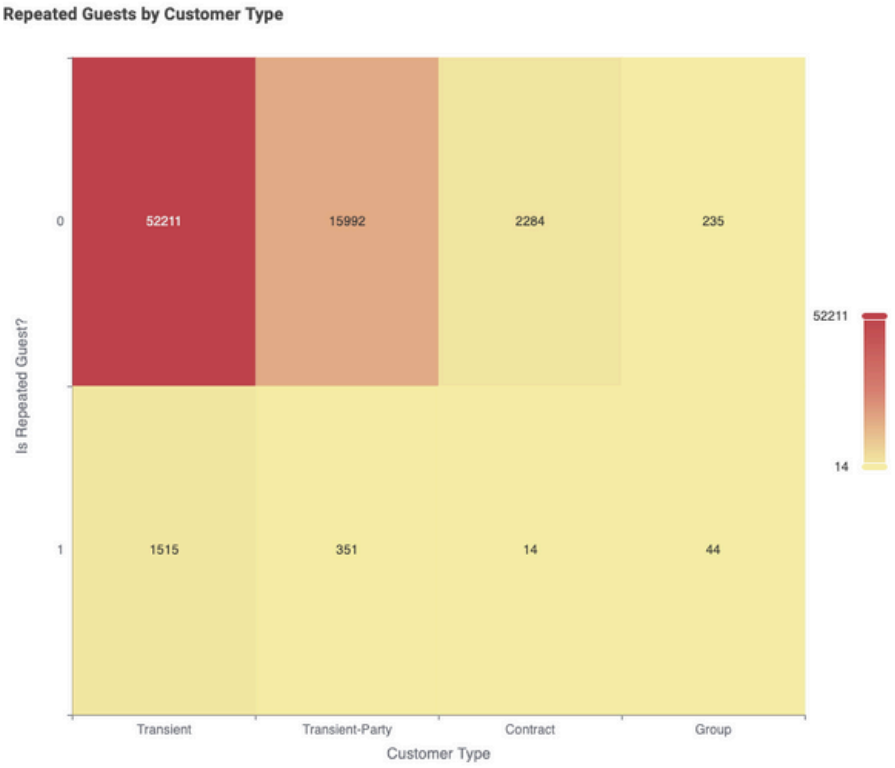
Rows: 26 | Columns: 14

Name	Type	# Missing val...	# Unique val...	Minimum	Maximum	25% Quantile	50% Quantile...	75% Quantile	Mean	Mean Absolu...	Standard Dev...	Sum	10 most com...
BookingID	Number (inte...	0	72646	1	72,646	18,161.75	36,323.5	54,485.25	36,323.5	18,161.5	20,971.238	2,638,756,981	1 (1; 0.0%), 2 ...
ArrivalDate	Local Date	0	731										2015-10-16 (...
LeadTime	Number (inte...	0	435	0	629	21	69	156	105.367	84.468	108.396	7,654,521	0 (2951; 4.06...
StaysInWeek...	Number (inte...	0	14	0	16	0	1	1	0.784	0.756	0.883	56,954	0 (35045; 48....
StaysInWeek...	Number (inte...	0	29	0	41	1	2	3	2.164	0.999	1.442	157,221	2 (24717; 34....
Adults	Number (inte...	0	5	0	4	2	2	2	1.838	0.366	0.507	133,525	2 (53261; 73....
Children	Number (dou...	4	4	0	3	0	0	0	0.084	0.157	0.357	6,070	0 (68381; 94....
Babies	Number (inte...	0	5	0	10	0	0	0	0.005	0.01	0.086	361	0 (72308; 99....
Meal	String	0	4										BB (57742; 7...
Country	String	24	160										PRT (29715; ...
MarketSegm...	String	0	8										Online TA (34...
DistributionC...	String	0	5										TA/TO (6313...
IsRepeatedG...	String	0	2										0 (70722; 97....
PreviousCanc...	Number (inte...	0	10	0	21	0	0	0	0.086	0.159	0.427	6,221	0 (67302; 92....
PreviousBook...	Number (inte...	0	68	0	67	0	0	0	0.128	0.252	1.616	9,327	0 (71167; 97....
ReservedRoo...	String	0	8										A (58050; 79....
BookingChan...	Number (inte...	0	21	0	21	0	0	0	0.179	0.314	0.599	13,023	0 (63606; 87....
DepositType	String	0	3										No Deposit (5...
Agent	Number (dou...	7686	220	1	495	9	9	19	27.663	30.267	55.12	1,797,001	9 (28144; 43....
Company	Number (dou...	69095	203	8	494	40	91	219	144.952	98.402	117.81	514,726	40 (856; 24.1...
DaysInWaitin...	Number (inte...	0	115	0	391	0	0	0	3.519	6.707	21.78	255,673	0 (69208; 95....
CustomerType	String	0	4										Transient (53...
ADR	Number (dou...	0	5086	0	5,400	77	97.54	122.4	102.825	28.643	42.794	7,469,793.76	62 (3593; 4.9...
RequiredCarP...	Number (inte...	0	4	0	3	0	0	0	0.025	0.048	0.155	1,781	0 (70870; 97....
TotalOfSpeci...	Number (inte...	0	6	0	5	0	0	1	0.525	0.651	0.77	38,124	0 (45041; 62....
IsCanceled	String	0	2										0 (41991; 57....

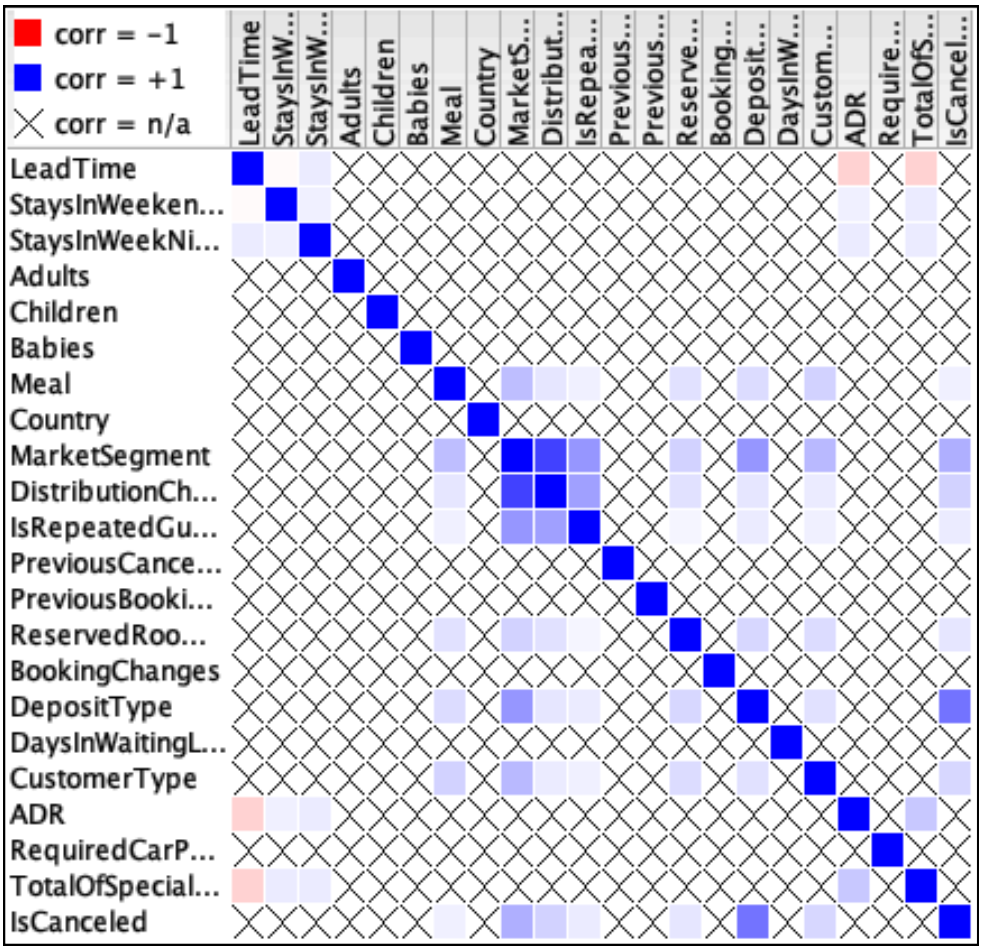
Appendix 1. Statistics View Table



Appendix 2. Bar Chart - Average Daily Rate by Customer Type



Appendix 3. Heatmap - Repeated Guests by Customer Type



Appendix 4. Correlation Matrix

		DT Pred	
		0	1
IsCanceled	0	10664	1933
	1	2197	7000

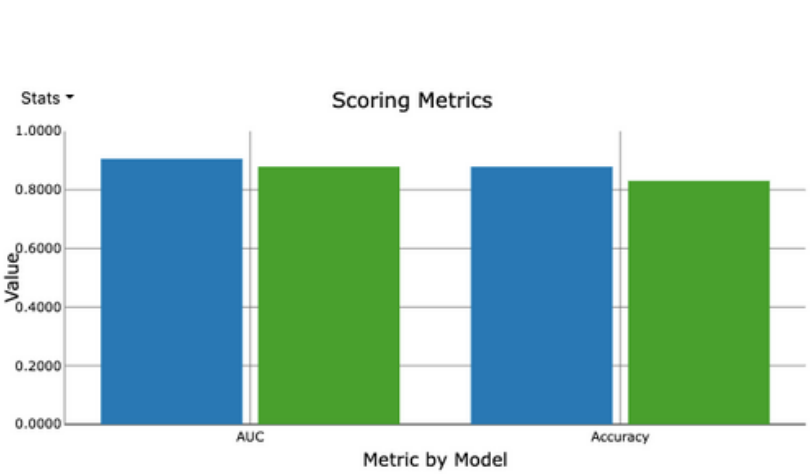
Appendix 5A. Confusion Matrix for Decision Tree

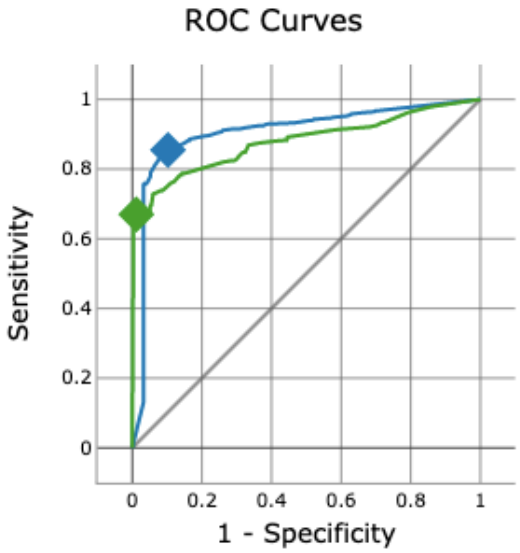
		RF Pred	
		0	1
IsCanceled	0	11842	755
	1	3469	5728

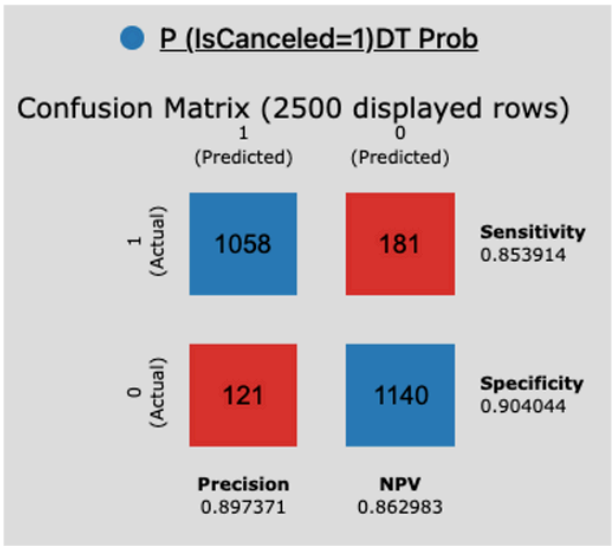
Appendix 5B. Confusion Matrix for Random Forest

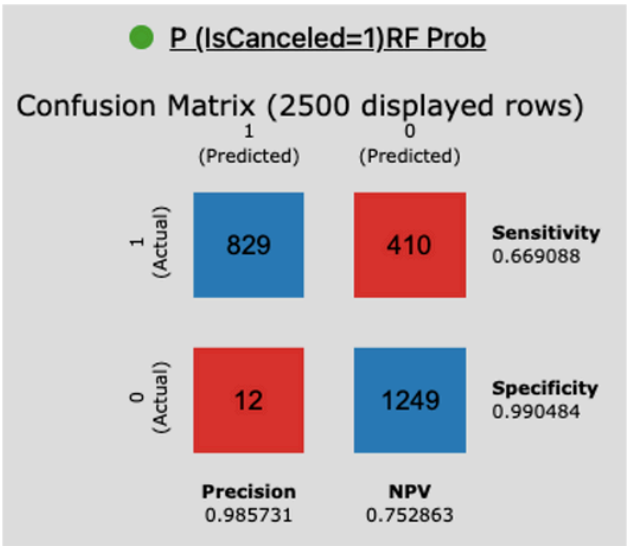
	Corrected Classified	Wrong Classified	Accuracy	Error
Decision Tree	17 664	4 130	81,05%	18,95%
Random Forest	17 570	4 224	80,619%	19,381%

Appendix 6. Scorers (Decision Tree and Random Forest)









Appendix 7. Binary Classification Inspector