LOREM IPSUM HOTEL GROUP

NOVA IMS Information Management School

2024/2025

Hotel Cancellation Prediction

Classification Report

We're ready to be part of
your spontaneous journey

Data Driven Decision Making Project

Anastasia Ciobanu
m20210516

Executive Summary

This report presents a data-driven approach to predicting hotel reservation cancellations using a dataset of more than 72.000 reservations. The main goal of this project is to create a machine learning model to help hotel managers anticipate cancellations, improving occupancy planning, staffing and revenue management. To structure and perform the work, the CRISP-DM framework and KNIME's low-code platform were used. Thus, the process was structured in terms of understanding the business, preparing the data, training the model and designing the implementation.

The Exploratory analysis revealed a cancellation rate of 42% and identified the key indicators, including *LeadTime*, *DepositType*, *CustomerType* and *TotalOfSpecialRequests*. During pre-processing, missing values were handled, categorical data was encoded, numerical features were normalized, and new variables were created, such as *TotalGuests* and *CancellationHistoryRate*.

The Decision Tree, Random Forest and XGBoost classifiers were evaluated. Due to its good ability to handle imbalanced data, XGBoost had the best performance (F1 score: 0,818) and provided actionable insights, for instance, high-risk bookings can be flagged for follow-up or policy adjustments, due to its good ability to handle unbalanced data.

In summary, this project demonstrates the how predictive modeling can provide tangible value to the hospitality industry by enabling proactive decision-making, minimizing revenue loss from cancellations and enhancing the customer experience through tailored interventions.

For illustration purposes, several visualisations and selected snippets of the model development workflow are included in the annex, offering a clear overview of key variables and methodological steps.

Business & Data Understanding

The data exploration phase was essential for identifying relevant patterns, inconsistencies and variables of interest in the dataset. The goal was to extract meaningful insights that could inform the model's ability to predict cancellations, while ensuring data quality, consistency and interpretability.

As mentioned before, the dataset contains over 72.000 rows representing individual hotel bookings from both a city and resort hotel and the target variable, *IsCanceled*, is a binary variable (0 = not cancelled, 1 = cancelled). A preliminary analysis using KNIME's Statistics, Box Plot and Value Counter nodes revealed a moderate class imbalance: approximately 42,2% of all bookings were cancelled. This needed careful evaluation metric selection and consideration of stratified sampling during model development.

On the other hand, several variables were identified as highly informative early in the analysis. *LeadTime* showed a strong positive correlation with cancellations — guests booking far in advance were more likely to cancel. The variable *DepositType* also played a important role - bookings with a non-refundable deposit had significantly lower cancellation rates than those with no deposit. Similarly, the number of *TotalOfSpecialRequests* and the customer classification (*CustomerType*) proved to be strong behavioral indicators.

As a result of this exploratory analysis, the following variables were identified for exclusion: *BookingID* - as it is a unique identifier with no predictive value -, *Company* - as it contained more than 95% missing values, making it statistically unreliable -, *DistributionChannel* as it was highly correlated with *MarketSegment* and was therefore considered redundant and, finally, *IsRepeatedGuest* - due to its overlap with *PreviousBookingsNotCanceled* and *PreviousCancellations*, making it statistically unreliable.

Data Engineering & Preparation

This section describes the steps implemented to make the data usable for modelling. First, the *ArrivalDate* variable, which was originally in string format, was converted into a datetime feature to be used for temporal analysis. From this variable, two new features were derived: the *MonthArrival* and the *DayWeekArrival*.

The variables *StaysInWeekendNights* and *StaysInWeekNights*, which separately recorded the number of nights a guest stayed during the weekend and the week, were aggregated into a single feature named *TotalNights*. From this, the ratio variable *IsWeekendStayRatio* was created, representing the proportion of weekend nights relative to the total number of nights stayed ($*StaysInWeekendNights*$ / $*TotalNights*$). These transformations simplified the stay duration representation and allowed the model to capture behavioral patterns associated with weekend travel, which might differ from weekday stays in terms of cancelation likelihood. Additionally, the variable *PreviousCancellations* was replaced by a more informative metric. The new feature determines whether there were any previous bookings and, if so, computes the proportion of those that were cancelled. If there were no previous bookings, the feature is set to zero. This created a more meaningful behavioral profile than the raw count.

The *RequiredCarParkingSpaces* binary variable was encoded into a categorical format with the values “Yes” and “No” to increase semantic clarity and signal intent to park rather than quantify space. Then, two similar new binary variables, *HadSpecialRequest* (set to “Yes” if the value is greater than 0 and “No” otherwise) and *WasInWaitingList* (indicating whether a booking has ever been on the waiting list) were derived from *TotalOfSpecialRequests* and *DaysInWaitingList*, respectively. On the other hand, a categorical variable name *CountryGroup* was created by grouping countries into five market categories: “Portugal”, “TopMarket”, “MediumMarket”, “LowMarket” and “MinorMarket”. This way, the high cardinality of the *Country* variable in reduced and the market-level analysis is simplified.

The categorisation was based on booking volume: countries with more than 3000 bookings were classified as “TopMarket”; those with 1001 to 3000 bookings as “MediumMarket”; countries with 501 to 1000 bookings fell under “LowMarket”; and all countries with fewer than 500 bookings were grouped into “MinorMarket”. Portugal was kept as a separate group given its central role in the dataset and business context.

Another categorical feature - *ArrivalSeason* - was introduced, mapping each month of arrival to its corresponding season (“Winter”, “Spring”, “Summer” or “Autumn”), allowing the model to take seasonality into account in cancellation behavior. The *Agent* variable, which contained both missing and valid values, was transformed into a binary feature indicating whether an agent was involved in the booking. As such, missing values were treated as “No” and existing values were labeled as “Yes”, as the specific agent identity was not relevant for the analysis.

To reduce noise, only relevant features were kept. As such, the variables *BookingID*, *ArrivalDate*, *Country* and *DaysInWaitingList* were removed due to redundancy and low variance. After this filtering, One-Hot-Encoding was applied to the remaining categorical features to ensure numerical representation and allow a uniform approach to subsequent preprocessing steps.

Following data engineering, the dataset was put through a comprehensive Data Preparation Phase to ensure quality and model interpretability. The data was first split into training and test sets, using an 80/20 ratio. Then, to address missing data, a mean imputation strategy was applied, replacing every missing entry with the mean of the corresponding non-null values. This approach was chosen for its simplicity and ability to maintain distribution integrity without introducing significant bias. Finally, outliers for numerical variables were detected and then replaced by the closest "normal" value.

Although the models used in later stages of the project were not distanced-based, a z-score normalization was performed to scale the features. This transformation standardizes all the features to a common scale, which allows for a better interpretability of the data and ensures uniformity between different preprocessing techniques. Finally, to examine the relationships, a linear correlation analysis was carried out. This helped identifying potential multicollinearity issues and provided insights on the dependencies of the features. In the Annex Section, a figure of the correlations is available for a better understanding.

Model Engineering & Evaluation

Three supervised machine learning algorithms—Decision Tree, Random Forest, and XGBoost—were trained and thoroughly assessed during the creation of this predictive system. In addition to comparing their performance using accepted evaluation metrics, the goal was to determine how well each model could adjust to the features of our dataset, specifically the existence of class imbalance and the need to reduce false positives and false negatives for business purposes.

Model	F1-Measure	Accuracy	Precision	AUC
Decision Tree	0,806	0,837	0,809	0,858
Random Forest	0,798	0,834	0,840	0,893
XGBoost	0,818	0,852	0,854	0,931

Table 1. Results of F1-Measure, Accuracy, Precision and AUC of each model

The modelling process began with the Decision Tree classifier, a foundational model known for its transparency and ease of interpretation. It gave a clear visual framework that demonstrated how various features affect classification results. But even though the Decision Tree was a useful baseline due to its simplicity, it showed limitations when used with unknown data. The accuracy and F1-score of the model were 0,837 and 0,806, respectively. The precision of 0,809 and AUC of 0,858 of the model, despite its initial promise, demonstrated its comparatively poorer capacity to strike a balance between sensitivity and specificity. This suggested an overfitting tendency, which would make it less dependable for use in practical situations where generalization is essential.

The Random Forest classifier was the tool we used to get around these limitations. The predictions of several decision trees are combined using this ensemble approach, which also introduces randomization into feature selection and data sampling. As a result, it reduced variance and improved model robustness. When tested, the Random Forest produced better accuracy at 0,834 and precision at 0,840, but its F1-score of 0,798 was marginally lower than the Decision Tree's. Notably, its AUC of 0,893 demonstrated a noticeably improved capacity to differentiate between classes at all decision thresholds. Better overall classification performance was shown by this increase in AUC, especially in detecting true positives and lowering the possibility of misclassification. Even though the Random Forest was more dependable and stable than the Decision Tree, it could still be improved, especially in terms of maximizing the F1-Score, which we gave priority to because of the dataset's moderate class imbalance.

In order to complete the model development process, XGBoost, a cutting-edge gradient boosting framework that has continuously surpassed other classifiers in structured data tasks, had to be implemented. A highly accurate and effective model is produced by XGBoost, which iteratively improves its predictions by concentrating on the errors made by prior learners. Using our dataset, XGBoost outperformed Random Forest and Decision Tree in every metric. The model's F1-score of 0,818 was the highest of all the algorithms that were tested. It achieved 0,852 accuracy, which shows a very good match between the predicted and actual labels, and 0,854 precision. Most notably, its AUC increased to 0,931, demonstrating XGBoost's superior capacity to accurately classify both positive and negative classes across a range of thresholds. The model was shown to be extremely sensitive and precise, making it the best option for deployment, as evidenced by its high AUC and best F1-score.

In addition to quantitative performance, XGBoost also provided useful qualitative insights. In line with previous conclusions from our exploratory data analysis phase, we were able to confirm the significance of important variables like *CustomerType* and *MarketSegment* by utilizing its integrated feature importance analysis. The model's conclusions gained additional credibility as a result of the coherence between exploratory insights and model interpretation.

All things considered, XGBoost was selected as the last model to be deployed following a rigorous engineering and assessment process. It showed a strong capacity for generalization, resilience to class disparity, and the capacity to produce reliable and understandable results. Its implementation is justified by its overall performance across all evaluation dimensions, which support the project's operational and strategic goals.

Deployment

The implemented model can be processed in batch and real-time, and it can be exported in PMML format using KNIME. The model has already been used to simulate batch scoring on unseen reservations on a held-out test set, validating this. Daily batch scoring could be used in the real world to assess recent reservations and identify possible cancellations, while a real-time API could offer immediate feedback at the time of booking confirmation. The final score would enable real-time booking status updates and set off focused operational reactions.

In terms of operations, the model makes strategic revenue optimization possible. It supports dynamic overbooking techniques on dates with a high risk. For reservations where cancellations are likely to occur (e.g. more than 70%), customized messages can be sent, flexible deposit policies can be implemented, or follow-up activities can be started. The model, on the other hand, permits the strategic upselling or distribution of premium services to visitors who are unlikely to cancel. Staffing and operational efficiency are also enhanced by cancellation forecasting. Increased customer segmentation, lower last-minute cancellations, higher occupancy rates, and better planning capabilities are all anticipated business impacts.

However, caution is required. Customer discontent or a loss of trust could result from false positives, and model accuracy could deteriorate if visitor behavior changes seasonally. A reliance on automation that is too great may also ignore unique circumstances and subtleties in visitor profiles.

Furthermore, adherence to the General Data Protection Regulation (GDPR) and other data protection laws is essential, particularly when it comes to automated processes. Combining human oversight with predictive outputs, keeping guest communications open, and regularly monitoring and retraining the model can all help to reduce these risks.

Monitoring & Maintenance

It is recommended to establish a monthly validation routine, comparing model predictions with actual booking results, which is crucial for recalibrating thresholds and monitoring key metrics (F1 score, precision and recall). KNIME's integrated capabilities, including workflow automation and the creation of dashboards, make this process efficient and scalable with minimal manual supervision.

To maintain a reliable feedback cycle, all forecasts and final results must be systematically recorded. These records will not only support performance monitoring but will also serve as training data for periodic retraining (which should be considered every three to six months, depending on the volume of bookings and behavioral volatility). A consistent discrepancy between predicted and actual results may indicate the need to update the model or its features.

To maintain a reliable feedback cycle, all forecasts and final results must be systematically recorded. These records will not only support performance monitoring but will also serve as training data for periodic retraining (which should be considered every three to six months, depending on the volume of bookings and behavioral volatility). A consistent discrepancy between predicted and actual results may indicate the need to update the model or its features. However, it is important to adapt to seasonal patterns and external events - such as vacation peaks or economic changes - which can alter cancellation trends.

To further improve the accuracy of the prediction, the hotel should consider collecting additional data not currently available (behavioral and technical). For example, variables such as the time spent on the booking site, the type of device, the sentiment of other customers regarding previous stays and the ability to respond to confirmation emails can reveal deeper behavioral signals that can influence the likelihood of cancellation.

As far as the model is concerned, although XGBoost has proven to be highly effective, exploring alternative models can offer complementary benefits or even better align with changing business needs. LightGBM, for example, offers similar predictive power with faster training and less use of resources. On the other hand, logistic regression, although less complex, offers transparency and ease of interpretation, something that is particularly useful in environments that require legal or operational clarity. More advanced approaches, such as stacked ensemble models or neural networks, may also be suitable if the data set is extended to include unstructured data, such as textual analysis or behavioral indicators. Testing different models in a controlled environment (e.g. A/B testing) can help determine whether a change in the algorithm would produce improvements in cancellation rate predictions.

Conclusion

The project emphasizes the iterative nature of data science workflows, where choices must continuously align with business goals, more specifically, this project shows how machine learning can deliver substantial value in the hospitality sector by converting raw booking data into actionable insights. Through a detailed process of data preparation, feature engineering and model evaluation, an accurate cancellation prediction model was developed to assist the hotel's operational needs.

One of the key findings is that guest cancellation behavior is shaped by a mix of behavioral, transactional and temporal factors, as variables such as *LeadTime*, *DepositType* and *TotalOfSpecialRequests* consistently emerged as strong predictors.

However, several challenges were encountered along the way. High-cardinality variables such as *Country* and *Agent* required careful grouping to avoid model noise and overfitting, while missing data on features such as *Company* led to their exclusion. Feature engineering, especially for behavioral metrics such as previous cancellations or weekend stay ratios, involved significant experimentation and a certain amount of intuition. In addition, the absence of richer behavioral data may have limited the model's full predictive potential.

Looking ahead, some improvements could be explored. One possibility would be to incorporate clustering in order to segment customers, which would allow the model to better capture the behaviors of different customer profiles, such as loyal customers versus those who are more price sensitive. Another addition would be the development of an interactive dashboard (using tools like Power BI or Tableau), to allow the easy visualization of cancellation risks and trends in real time. Finally, A/B tests could be implemented to help assess whether adjustments to the algorithm lead to measurable improvements in predicting cancellation rates.

In conclusion, this project not only delivered a high-performing predictive model but also laid the groundwork for a more data-driven and forward-thinking approach to hotel operations.

Annexes

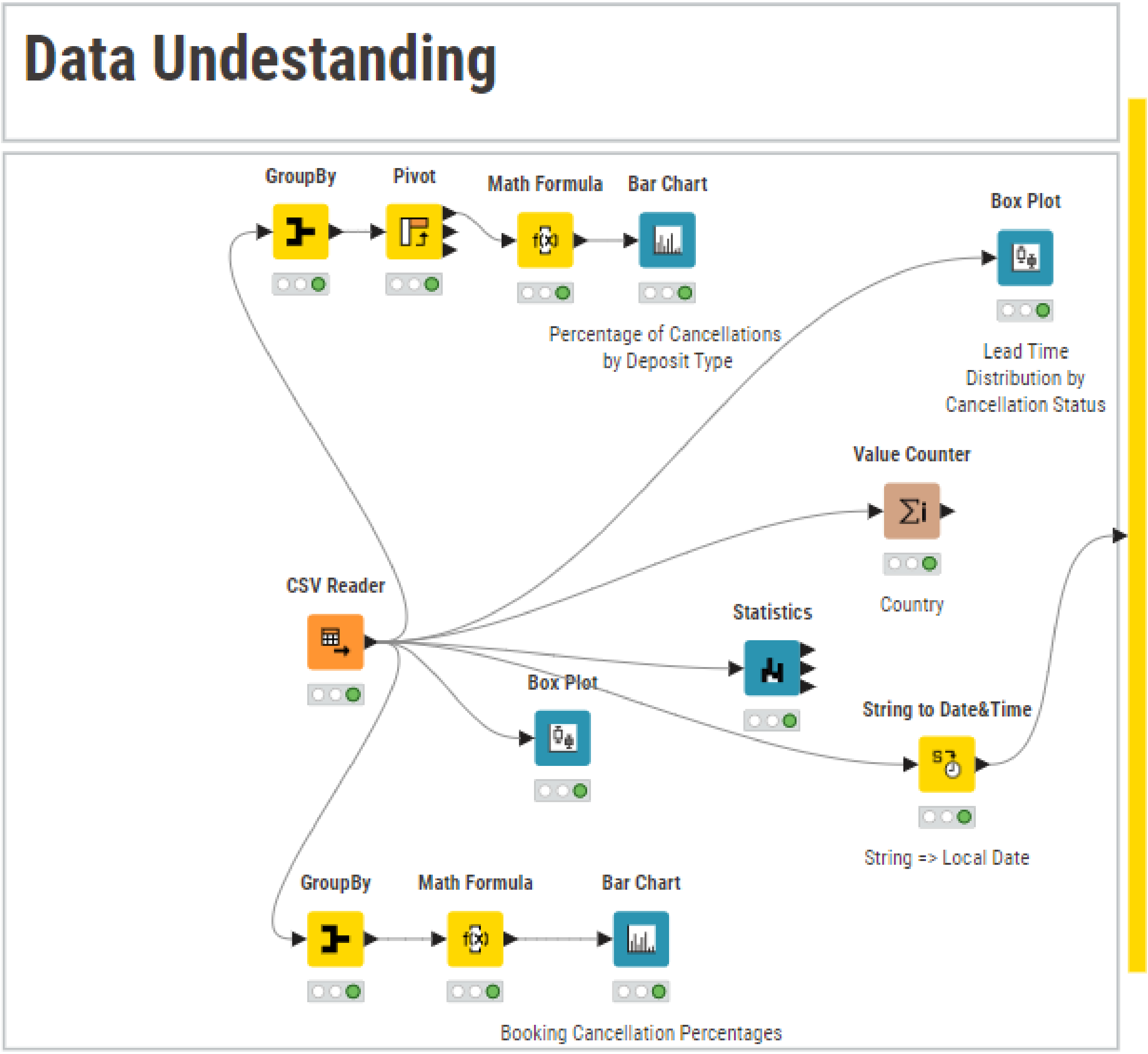


Fig.1 Booking Cancellations by percentages

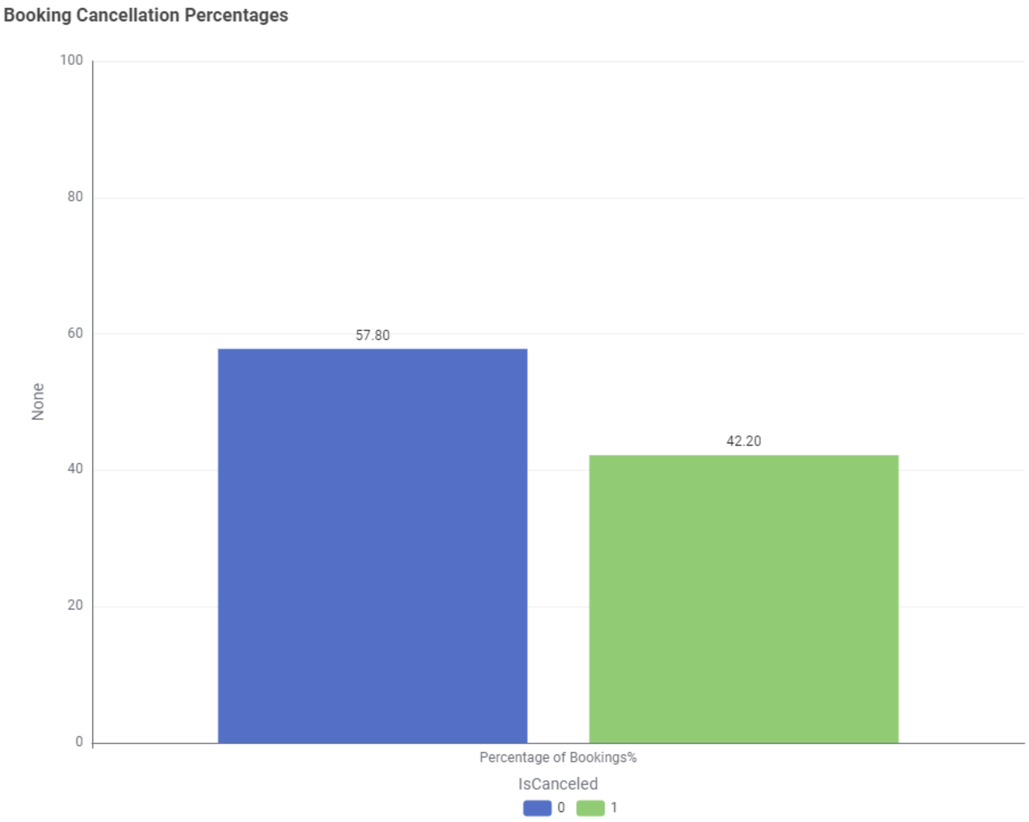


Fig.2 Percentage of Booking cancellations

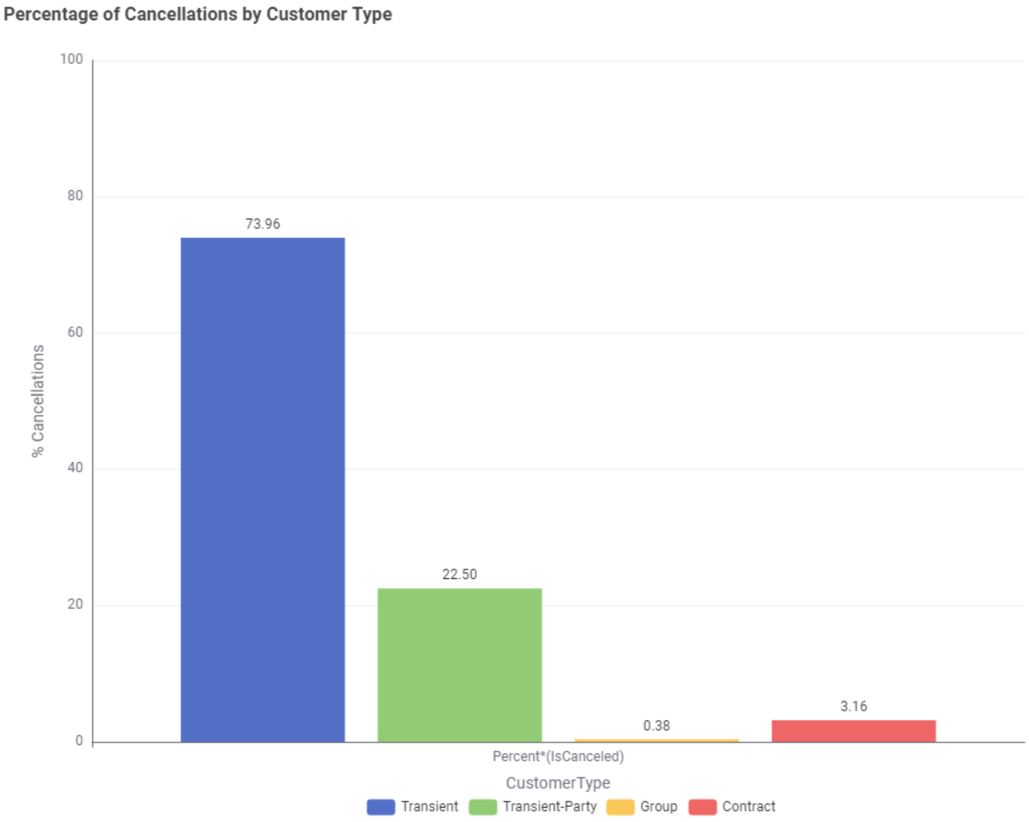


Fig.3 Percentage of Cancellations by Customer Type

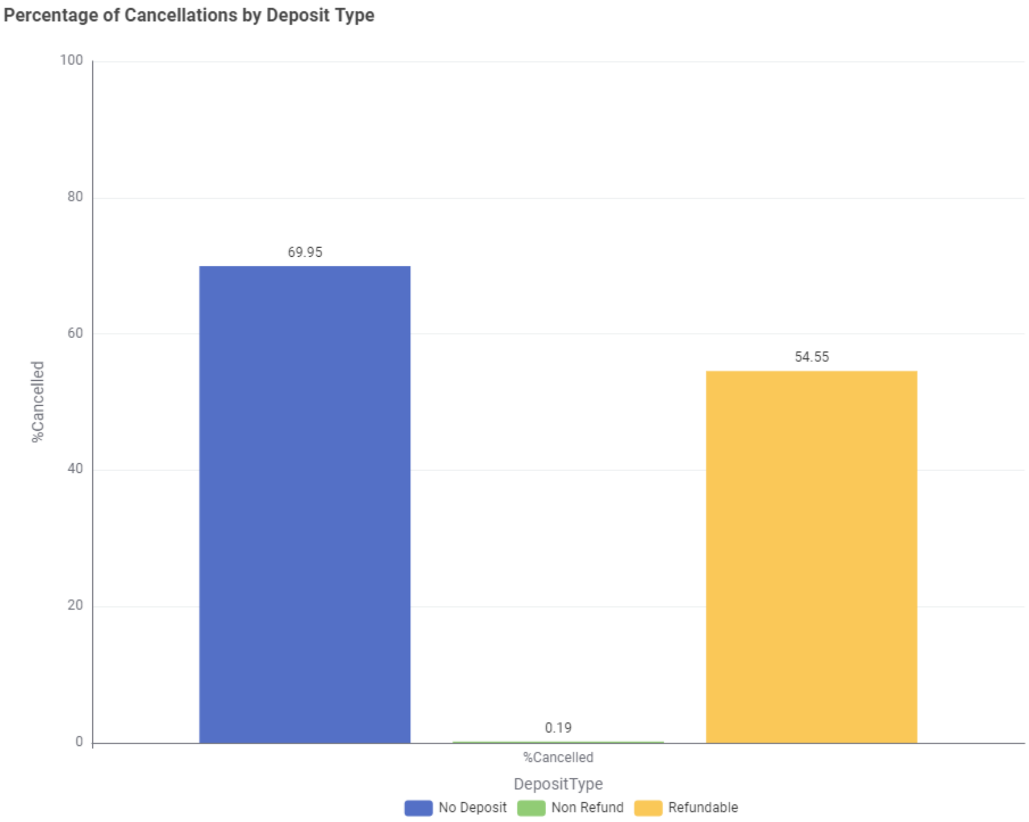


Fig.4 Percentage of Cancellations by Deposit Type

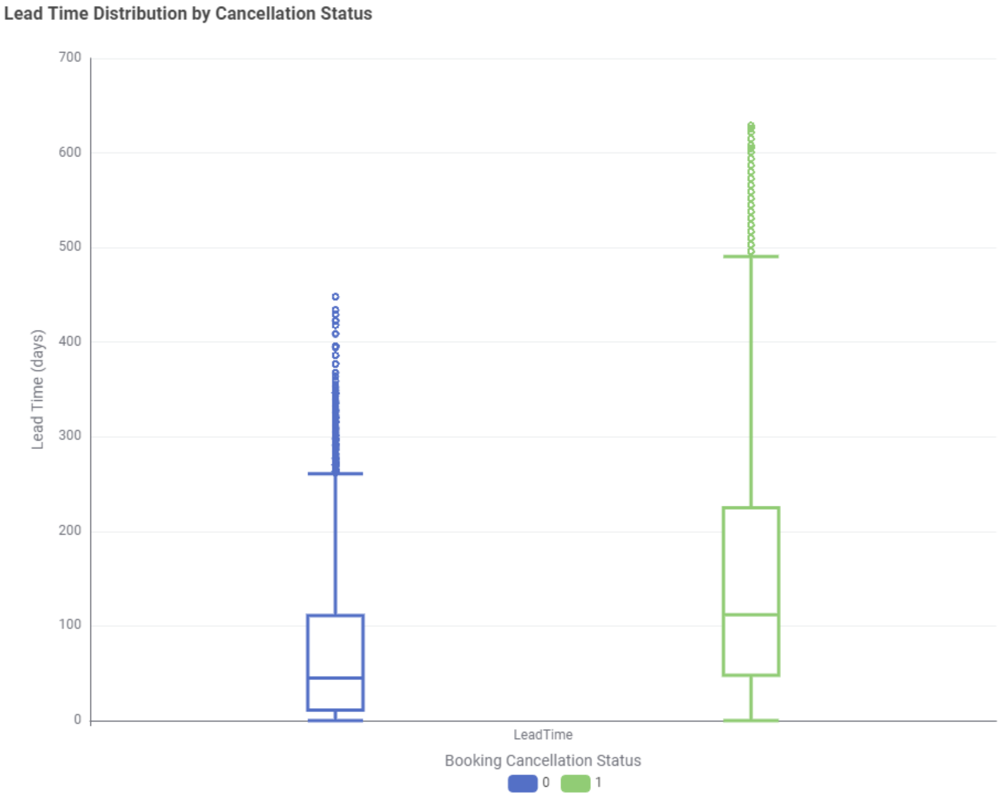
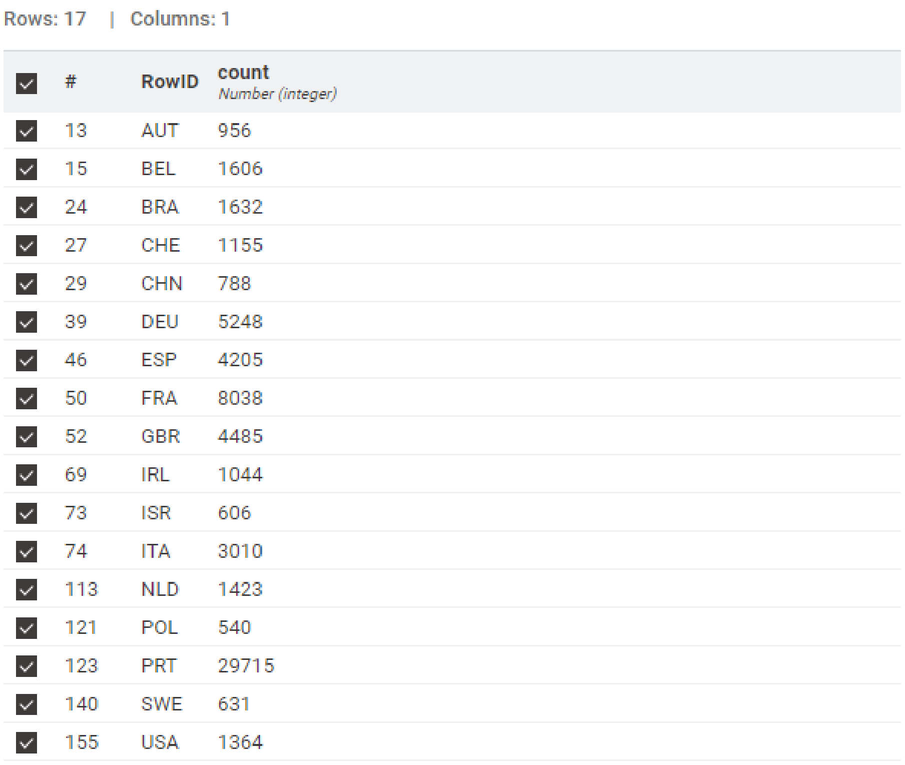


Fig. 5 Box plot of Lead Time Distributions by Cancellation Status



Rows: 17 | Columns: 1

☑	#	RowID	count *Number (integer)*
☑	13	AUT	956
☑	15	BEL	1606
☑	24	BRA	1632
☑	27	CHE	1155
☑	29	CHN	788
☑	39	DEU	5248
☑	46	ESP	4205
☑	50	FRA	8038
☑	52	GBR	4485
☑	69	IRL	1044
☑	73	ISR	606
☑	74	ITA	3010
☑	113	NLD	1423
☑	121	POL	540
☑	123	PRT	29715
☑	140	SWE	631
☑	155	USA	1364

Fig. 6 Selected Countries for the variable Country Group

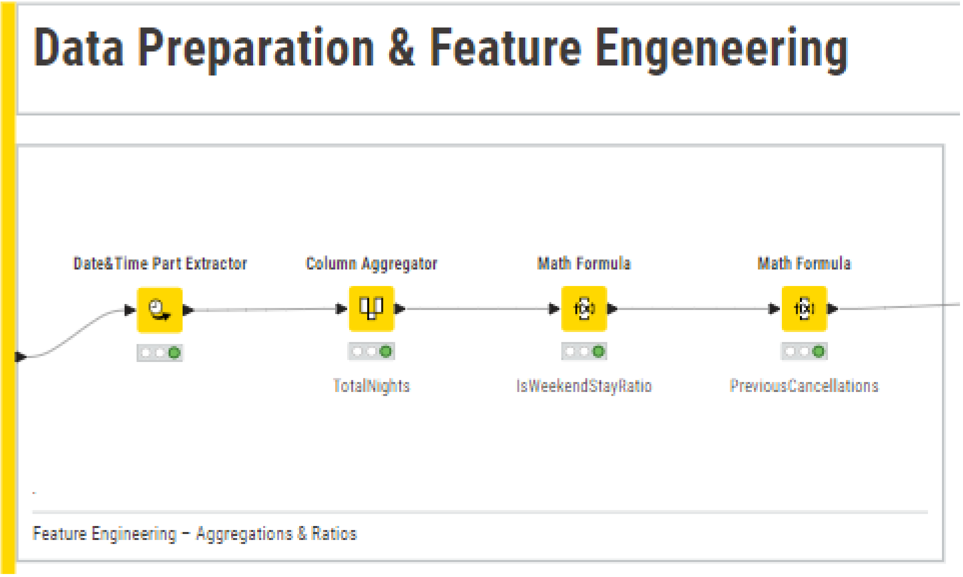


Fig.7 Data Preparation section - Aggregations and Ratios

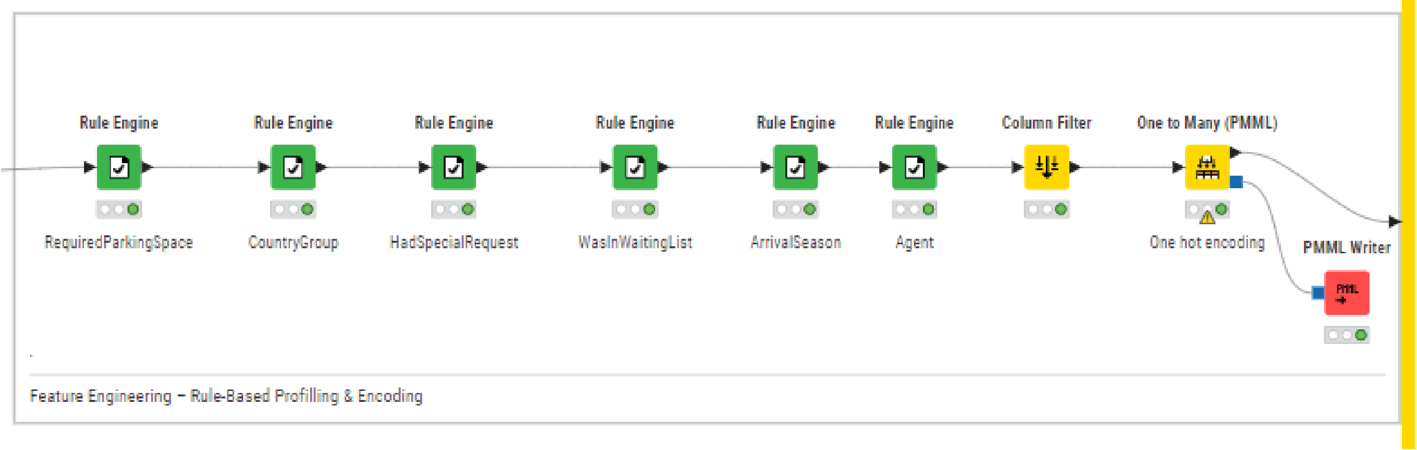
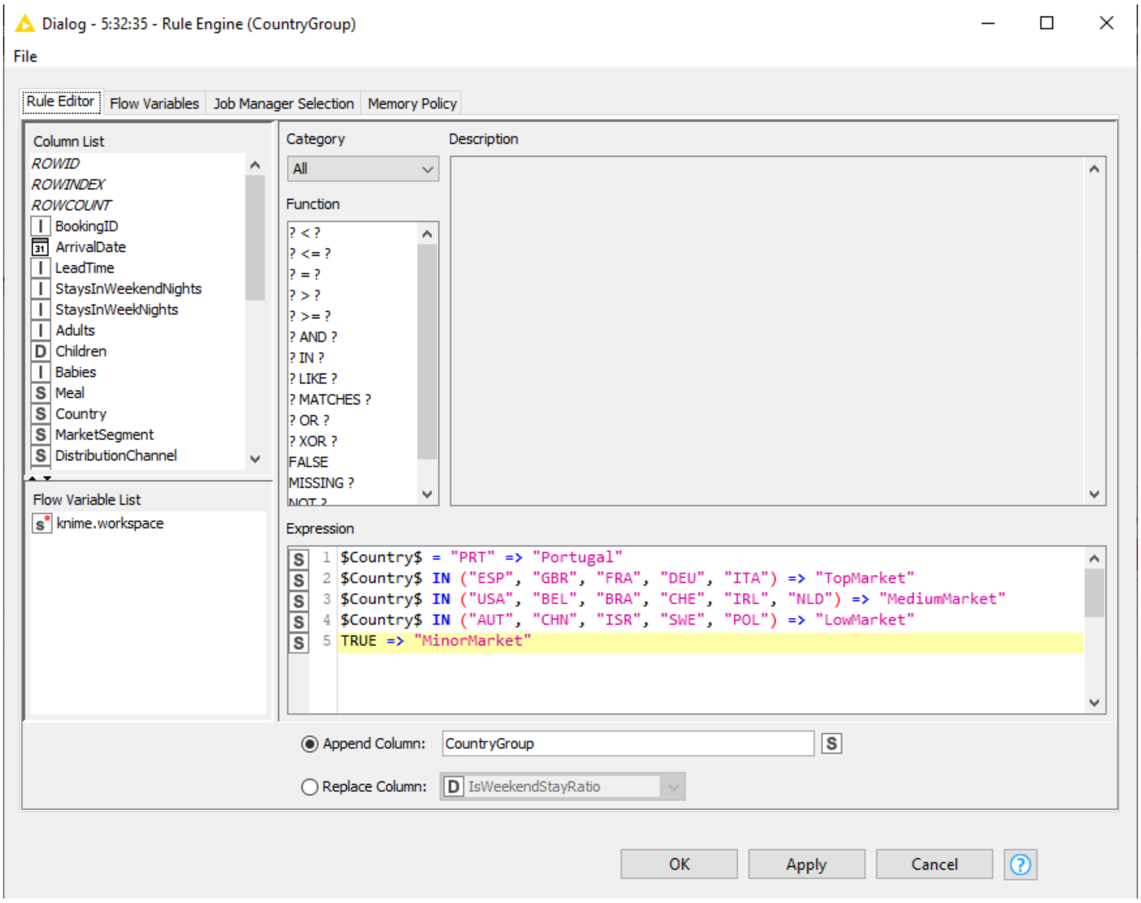


Fig. 8 Data Preparation and Feature Engineering- Rule-based profiling and encoding



```
$Country$ = "PRT" => "Portugal"
$Country$ IN ("ESP", "GBR", "FRA", "DEU", "ITA") => "TopMarket"
$Country$ IN ("USA", "BEL", "BRA", "CHE", "IRL", "NLD") => "MediumMarket"
$Country$ IN ("AUT", "CHN", "ISR", "SWE", "POL") => "LowMarket"
TRUE => "MinorMarket"
```

● Append Column: CountryGroup

○ Replace Column: IsWeekendStayRatio

Fig.9 Rule engine node of the *CountryGroup* variable

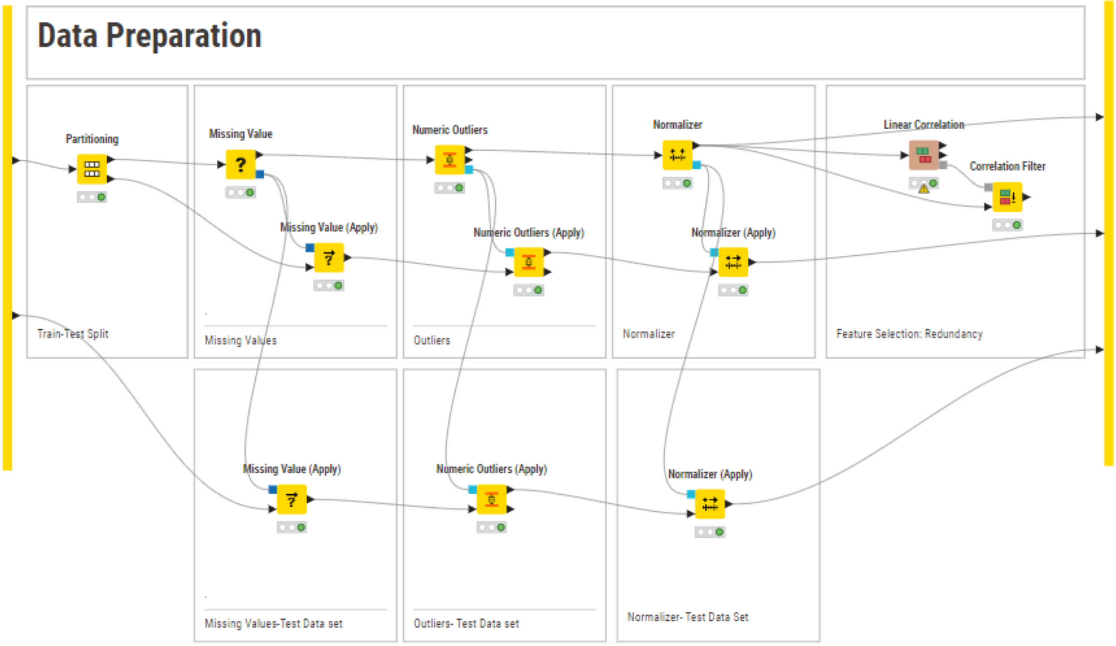


Fig. 10 Data Preparation Section Nodes

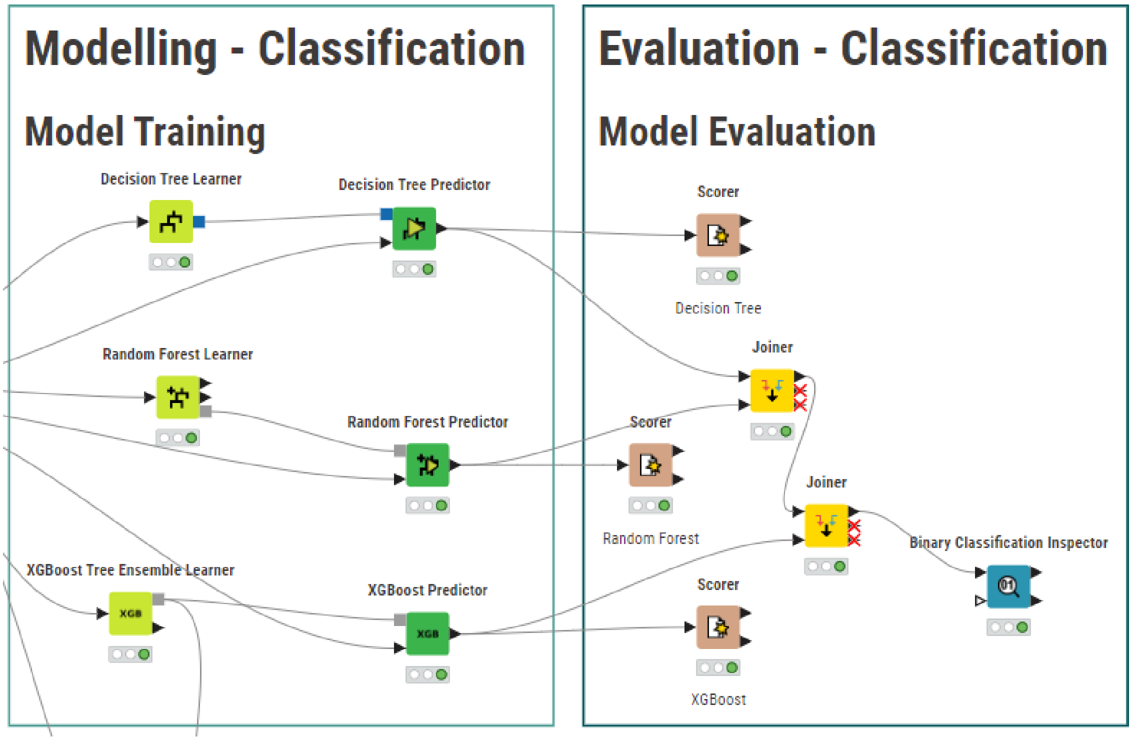


Fig.11 Modelling and Evaluation Section

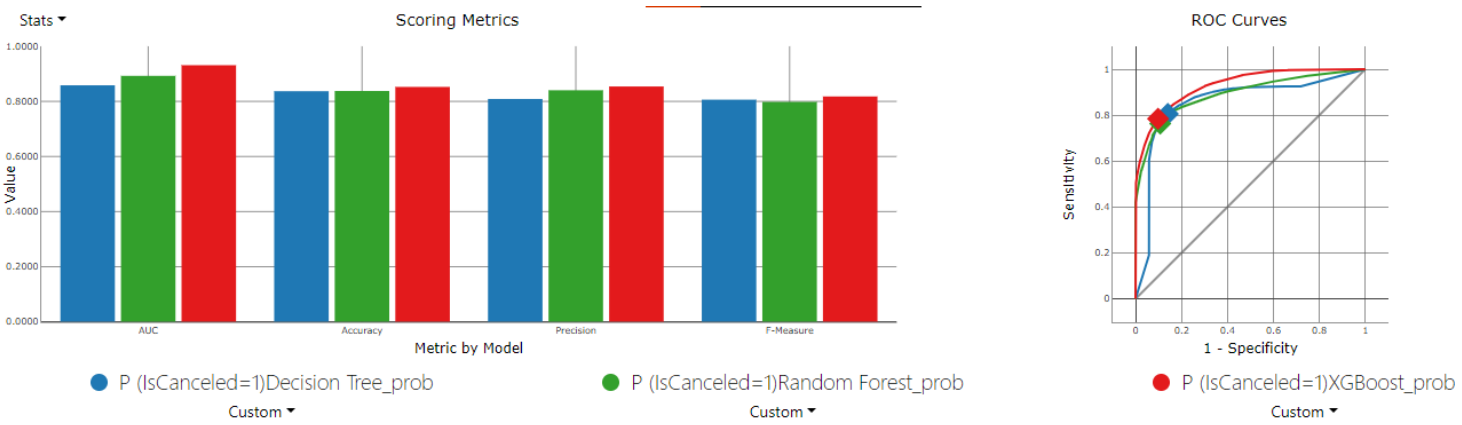
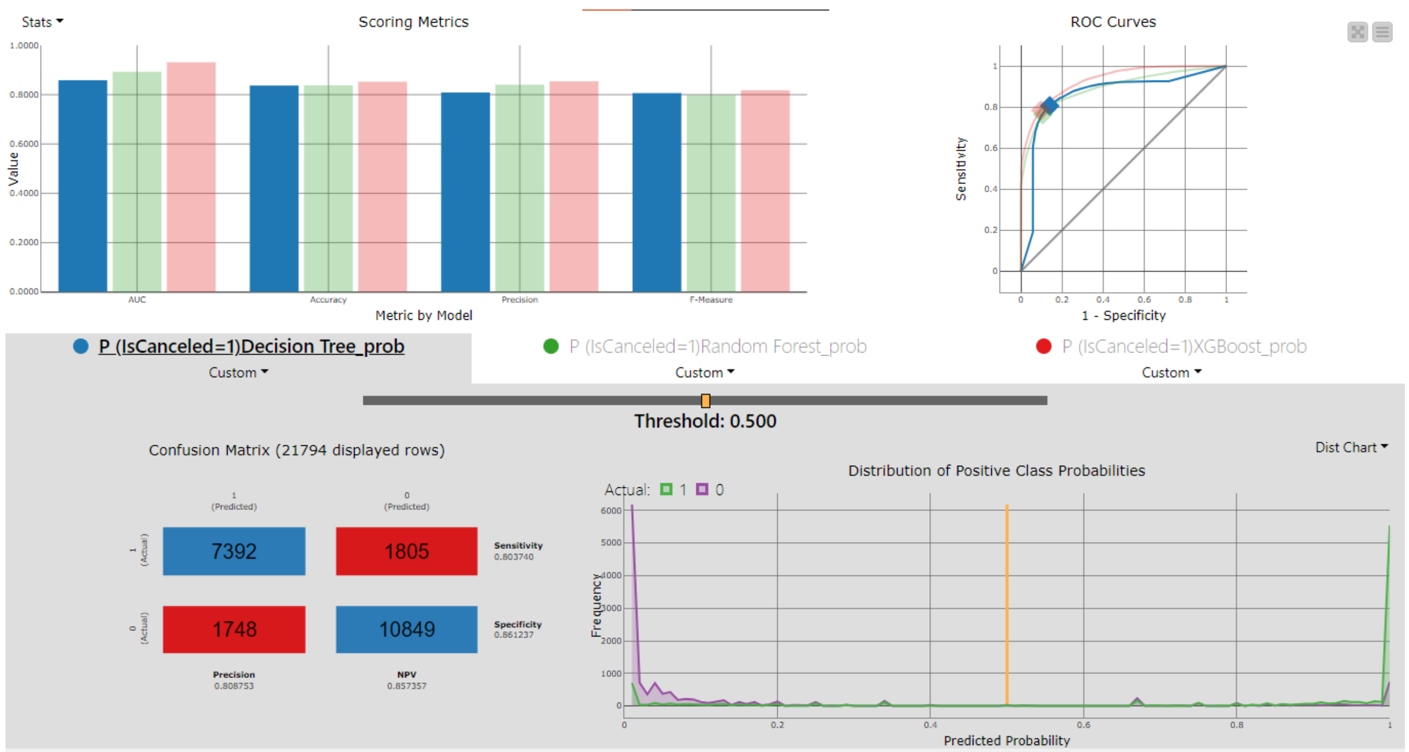


Fig. 12 Binary Classification Inspector of each model

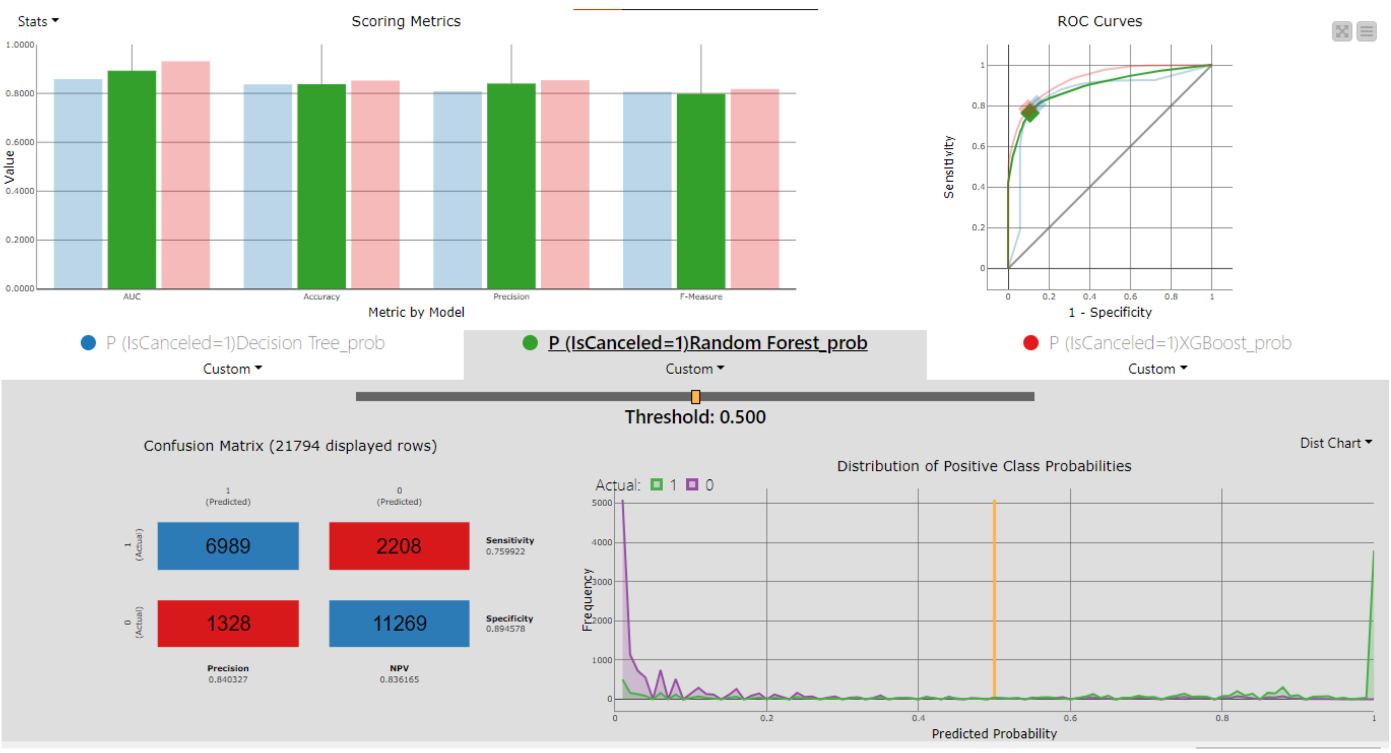


Threshold: 0.500

Confusion Matrix (21794 displayed rows)

	1 (Predicted)	0 (Predicted)	
1 (Actual)	7392	1805	**Sensitivity** 0.803740
0 (Actual)	1748	10849	**Specificity** 0.861237
	Precision 0.808753	**NPV** 0.857357	

Fig.13 Binary Classification for the Decision Tree Model

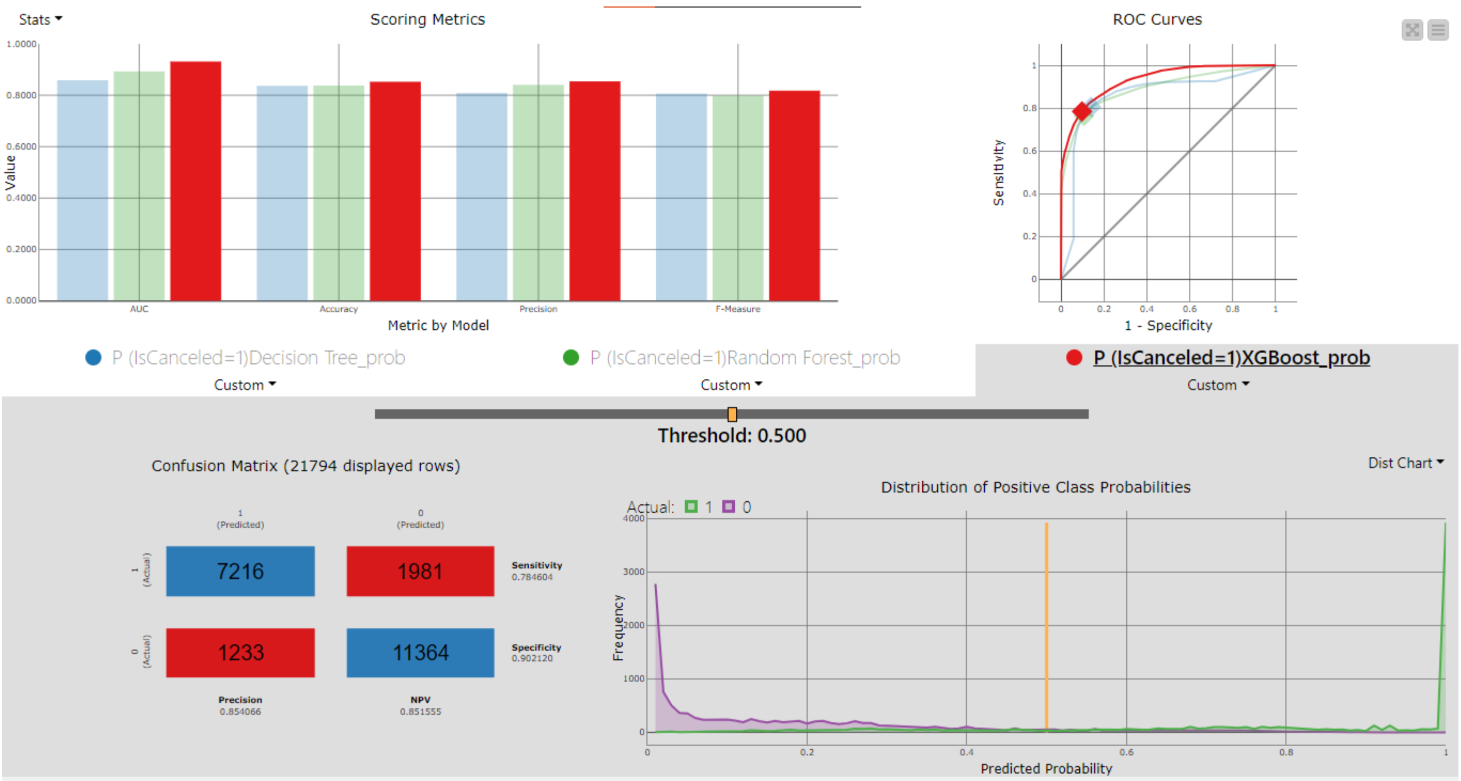


Threshold: 0.500

Confusion Matrix (21794 displayed rows)

	1 (Predicted)	0 (Predicted)	
1 (Actual)	6989	2208	Sensitivity 0.759922
0 (Actual)	1328	11269	Specificity 0.894578
	Precision 0.840327	NPV 0.836165	

Fig.14 Binary Classification for the Random Forest Model



Threshold: 0.500

Confusion Matrix (21794 displayed rows)

	1 (Predicted)	0 (Predicted)	
1 (Actual)	7216	1981	Sensitivity 0.784604
0 (Actual)	1233	11364	Specificity 0.902120
	Precision 0.854066	NPV 0.851555	

Fig.15 Binary Classification for the XGBoost Model

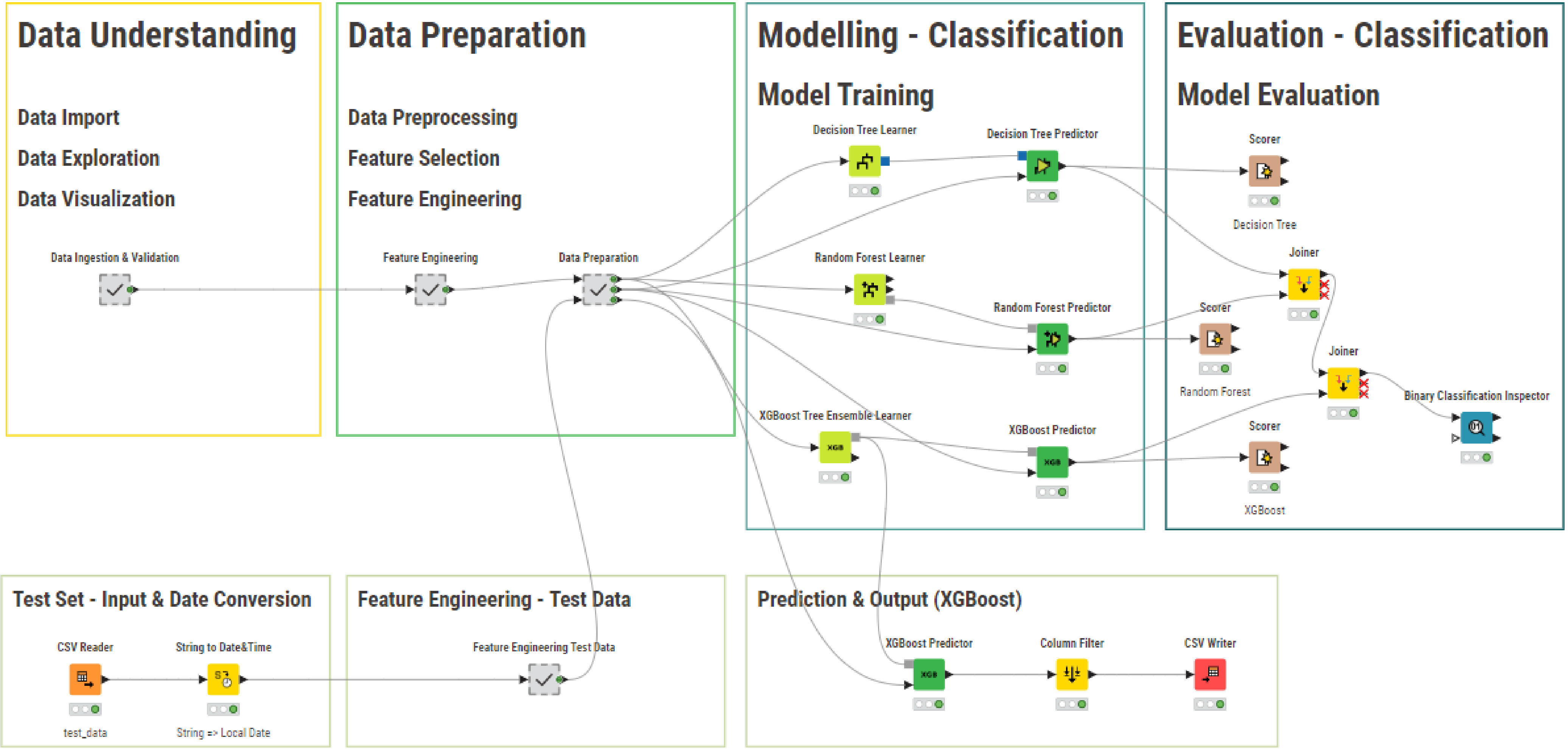


Fig.16 Overview of the KNIME Workflow for Hotel Cancellation Prediction