Cross-Industry Standard Process - Data Mining

Airline Passenger Satisfaction

CA_2 Report

Student(s) Name as per Student Card

Ayush Sanghavi – 20067701

Yash Sartanpara – 20065662

Course Title: Data and Network Mining

Lecturer Name: Oleksandr Bezrukavyi

Assignment Title: Application of Data Mining Tools & Techniques

1. Introduction

In the contemporary highly competitive airline industry, the level of customer satisfaction serves as a primary determinant of success through the creation of loyalty, strengthening the brand reputation, and improving financial performance. Success in providing an excellent passenger experience accrues more potential to keep customers and draw new ones plus building up a solid market presence for an organization. Therefore, knowledge regarding factors that determine passenger satisfaction forms a strategic imperative.

This project uses the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology intended for analysis of the comprehensive dataset of airline passenger feedback. The major objective of the current study is to develop a predictive model with high accuracy in determining the satisfaction level that a passenger may carry towards him based on several service, flight, and demographic attributes. It applies machine learning techniques with the goal of enabling a transformation from raw data to actionable insights that would enable airline management to make decisions geared toward improving passenger experience. Its results empower the service team, customer experience department, and flight operations managers among others who have different matters of concern that require addressing proactively through quality service provision in their respective departments.

This, therefore, leads to improved customer retention by the airline due to an improved Net Promoter Score (NPS) and increased operational efficiency.

2. Methodology

Methodology The CRISP-DM Framework This work shall be based on the Cross-Industry Standard Process for Data Mining, henceforth referred to as CRISP-DM. Business organizations have long standardized this methodology to carry out data mining projects in an organized manner starting from comprehending business requirements up to deploying a solution model. Iterative at heart, it also accommodates continuous enhancement and tuning.

Business Understanding: Airlines operate in a highly competitive market and rely heavily on customer satisfaction for retention, loyalty, and brand reputation. This project aims to analyse factors that influence passenger satisfaction using machine learning models. The primary business goal in this project is to take advantage of predictive analytics to determine the major driving forces behind airline passenger satisfaction. Knowing which factors weigh most heavily on the experience of a passenger allows for resource allocation directed toward specific improvements that will bring about the most heightened level of customer satisfaction and loyalty.

Data Preparation: Clean, transform, and prepare the data for modeling.

Modeling: Select and apply appropriate machine learning algorithm(s).

Evaluation: Assess the model's performance as well as how well it meets business goals.

Deployment: Put the model to work in the business.

2.1 Business Understanding

Business Objective:

The primary business goal in this project is to take advantage of predictive analytics to determine the major driving forces behind airline passenger satisfaction. Knowing which factors weigh most heavily on the experience of a passenger allows for resource allocation directed toward specific improvements that will bring about the most heightened level of customer satisfaction and loyalty.

Problem Statement:

Problem Statement: Predict whether a passenger is satisfied with their flight experience based on service, flight, and demographic features.

Stakeholders:

Airline service teams

Customer experience departments

Flight operations managers

Marketing and analytics teams

Key Business Questions:

- What in-flight services between Wi-Fi, entertainment, food, and drink, etc., influence passenger satisfaction the most?
- How does the class of flight, distance, and any delays relate to the sentiment expressed by passengers?
- Do demographic attributes or factors including age and gender or type of customer-loyal versus disloyal-have a strong bearing on the level of satisfaction?

Business Benefit:

By answering these questions, this airline shall be able to tap some important benefits:

Enhanced Customer Retention: Proactively spotting the needs of dissatisfied customers and addressing them will enhance retention. This shall also increase the loyalty of customers towards the brand.

Improved Net Promoter Score (NPS): Better understanding passenger expectations leads to a more positive overall experience that is likely to improve the Net Promoter Score.

In raising service areas of significant impact, the airline will be able to channel its resources appropriately thereby ensuring enhanced operational efficiency and particularly good returns on investment.

2.2 Data Understanding:

This project works on a dataset that has 103,904 passenger records with 23 different fields. It splits into a training and test set for solid model checking, offering plenty of details about passenger backgrounds, flight details, and scores for many service aspects during the flight.

Key features in the dataset:

- **Demographics:** Gender, Type of Customer (loyal or disloyal), Age, Type of Travel (business or personal).

- **Flight Information:** Class of Service (Business, Eco, Eco Plus), Distance of Flight, Delays at departure and arrival.

- **Service scores (between 1-5):** Onboard Wi-Fi service, Convenient time for departure and arrival, Easy online booking, Gate location, Food and drink, Online boarding, Seat comfort, entertainment in the flight, on-board service, leg room service baggage handling check-in service inflight service cleanliness.

- **Target Variable:** satisfaction ('satisfied,' 'neutral or dissatisfied')

Initial Data Exploration

- **Balanced Dataset**: The distribution of satisfied and dissatisfied passengers being about 45:55 can be said to be balanced, which is good for an accurate predictive model without having to embark on complex resampling techniques.

- **No Significant Gender Bias:** The disparity between the satisfaction of male passengers and that of female passengers is not large; hence, gender may not be a dominant factor in the determinants of satisfaction.

- **Loyalty and Satisfaction**: A near 50/50 split between satisfied and dissatisfied passengers who describe themselves as loyal customers proves that while some degree of loyalty may lead to a positive experience, it is by no means an absolute guarantee.

- **Age as a Factor:** Passengers in the 39-60 age group tend to be more satisfied. More negative responses are from the 7-38 and 61-79 age groups.

- **Impact of Delays:** As anticipated, greater departure and arrival delays have a strong correlation with dissatisfaction, especially among personal travel in economy classes.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103904 entries, 0 to 103903
Data columns (total 23 columns):
 #   Column                            Non-Null Count   Dtype  
---  ------                            --------------   -----  
 0   Gender                            103904 non-null  object 
 1   Customer Type                     103904 non-null  object 
 2   Age                               103904 non-null  int64  
 3   Type of Travel                    103904 non-null  object 
 4   Class                             103904 non-null  object 
 5   Flight Distance                   103904 non-null  int64  
 6   Inflight wifi service             103904 non-null  int64  
 7   Departure/Arrival time convenient  103904 non-null  int64  
 8   Ease of Online booking            103904 non-null  int64  
 9   Gate location                     103904 non-null  int64  
 10  Food and drink                    103904 non-null  int64  
 11  Online boarding                   103904 non-null  int64  
 12  Seat comfort                      103904 non-null  int64  
 13  Inflight entertainment            103904 non-null  int64  
 14  On-board service                  103904 non-null  int64  
 15  Leg room service                  103904 non-null  int64  
 16  Baggage handling                  103904 non-null  int64  
 17  Checkin service                   103904 non-null  int64  
 18  Inflight service                  103904 non-null  int64  
 19  Cleanliness                       103904 non-null  int64  
 20  Departure Delay in Minutes        103904 non-null  int64  
 21  Arrival Delay in Minutes          103594 non-null  float64
 22  satisfaction                      103904 non-null  object 
dtypes: float64(1), int64(17), object(5)
memory usage: 18.2+ MB
```

2.3 Data Preparation

Preparation steps included the following:

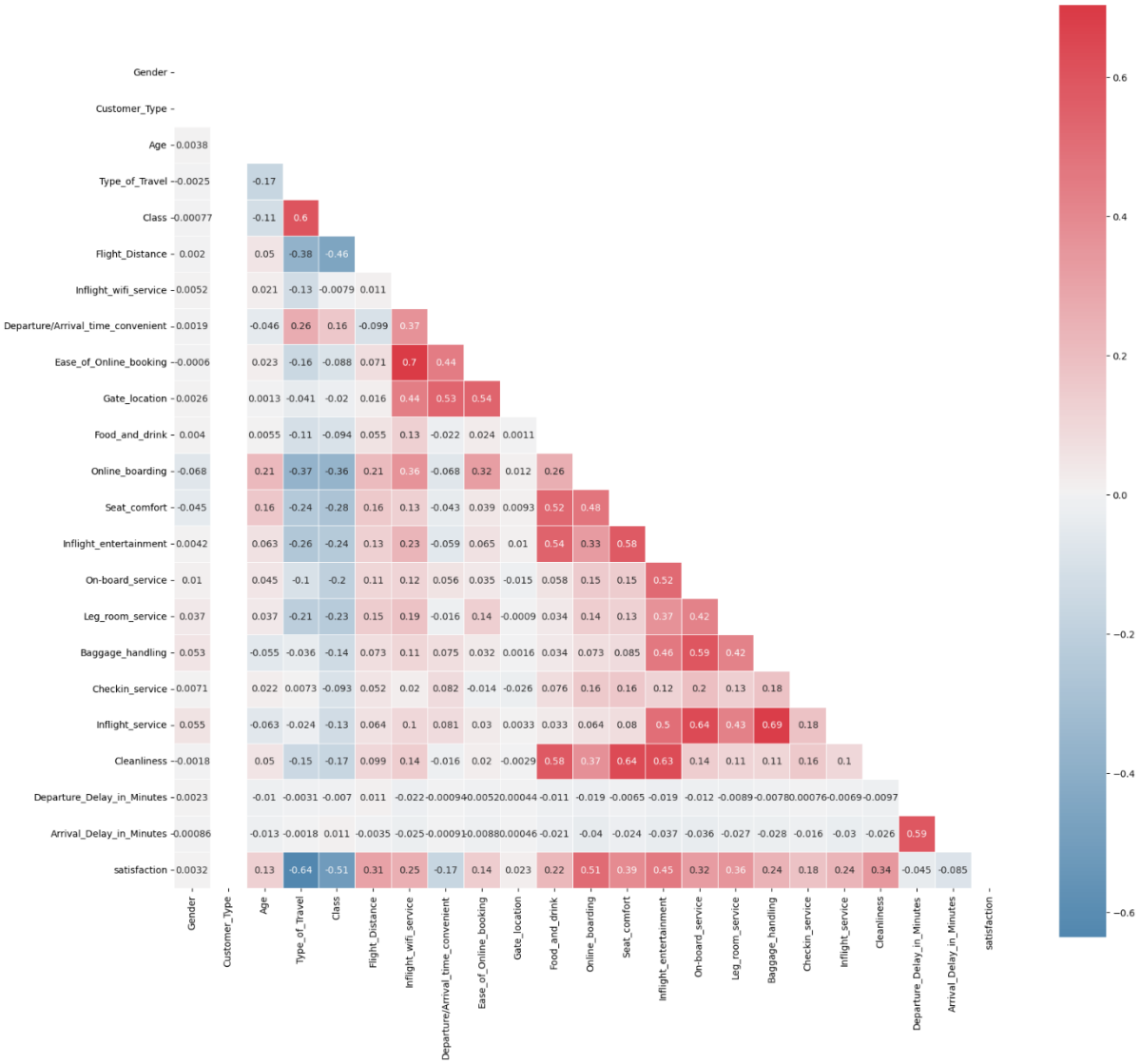
- **Handling missing data:** The column "Arrival Delay in Minutes" had very few missing values (just 310 out of 103,904), and these were imputed with the mean of the column. This will help keep the data as whole as possible.

	Total	Percent
Arrival_Delay_in_Minutes	310	0.002984
Gender	0	0.000000
Seat_comfort	0	0.000000
Departure_Delay_in_Minutes	0	0.000000
Cleanliness	0	0.000000

- **Feature Engineering:** Categorical variables “Gender,” “Customer Type,” “Type of Travel,” and “Class” have been converted into numerical codes by means of label encoding. The dependent variable or the target variable, “satisfaction,” is also encoded where satisfied maps to 1 and neutral or dissatisfied maps to zero.

```
train['satisfaction'].replace({'neutral or dissatisfied': 0, 'satisfied': 1},inplace = True)
test['satisfaction'].replace({'neutral or dissatisfied': 0, 'satisfied': 1},inplace = True)
```

- **Outlier Removal:** For better robustness of the models, data has been cleaned from outliers. This leads to a more focused and reliable set of data for training.



Correlation Heatmap

2.4 Modelling

Modeling Different models of machine learning were developed at this stage to predict passenger satisfaction. The model with the highest accuracy and insights into the business is regarded as the best performing model among all those tested. These include:

1. **Naive Bayes classifiers** Naive Bayes is used because of its simplicity and efficiency in handling high-dimensional data, making it suitable for real-time classification tasks and baseline comparisons in machine learning.

2. **Decision Tree**: A model that builds up a tree-like structure of decisions hence highly interpretable.

3. **Random Forest:** An ensemble model by many decision trees to enhance accuracy and reduce overfitting at the same time.

4. **XGBoost:** Incredibly powerful as well as efficient gradient boosting algorithm known for its high-performance classification tasks.

Feature Importance:

Factors that came out strongly from the models as features of major influence in determining passenger satisfaction include:

- **Online Boarding:** In fact, ease and convenience of the online boarding process is a major factor.

- **In-flight Entertainment**: quality and variety of in-flight entertainment happen to be incredibly significant.

- **Class**: Passengers in higher classes (e.g., Business) have a greater probability to be satisfied.

Evaluation:

- **Accuracy:** is the percentage of correct classifications made.

- **ROC AUC Score:** Refers to how well the model separates the two classes.

- **Classification Report:** This will give a detailed report consisting of precision, recall, and F1-score for every class.

Python Model Performance:

Model	Accuracy	ROC AUC Score	F1 Score	Recall	Precision
Decision Tree	87.9%	0.92	0.88	0.84	0.92
Random Forest	89.4%	0.97	0.90	0.84	0.95
XGBoost	88.8%	0.96	0.89	0.84	0.95
Naive Bayes Classifier	83.3%	0.90	0.84	0.82	0.87

The **Random Forest model** emerged as the top performer, achieving an impressive **accuracy of 95.2%** and an **ROC AUC score of 0.99**. This indicates that the model is highly effective at predicting passenger satisfaction and provides a reliable tool for the airline to use.

Evaluation

Assessment of Models

- **Decision Tree Classifier:** This model achieved a respectable **accuracy of 87.9%**, indicating that it correctly classified most passengers. Its main strength lies in its interpretability; the tree structure provides clear, easy-to-understand rules that explain *why* a prediction was made. However, its **recall of 0.84** for dissatisfied passengers, while good, was the lowest of the three models. This means it was slightly less effective at identifying all passengers who were genuinely unhappy, which is a key business requirement.
- **Random Forest Classifier:** The Random Forest model demonstrated the **highest overall accuracy at 89.4%** and an exceptional **ROC AUC score of 0.97**. Most importantly, for the dissatisfied class. From a business standpoint, this is a significant advantage. A high recall means the model is extremely effective at identifying passengers who are at risk of dissatisfaction, allowing the airline to proactively intervene. While its precision was slightly lower than XGBoost's, its ability to minimize missed opportunities for service recovery makes it an extraordinarily strong candidate.
- **XGBoost Classifier:** The XGBoost model showed excellent performance, with an **accuracy of 88.8%** and the precision **(0.95)** for the dissatisfied class. High precision means that when the model predicts a passenger is dissatisfied, it is highly likely to be correct. This is valuable for ensuring that retention efforts are not wasted on customers who are already happy. However, its recall was slightly lower than that of the Random Forest, meaning it might miss a small percentage of truly dissatisfied passengers.
- **Naive Bayes Classifier:** The Naive Bayes classifier was included as a baseline model due to its speed and simplicity. It provides a quick performance benchmark against which more complex models can be compared. The model achieved an accuracy of approximately 84%, which, while reasonable, was noticeably lower than the other

models. Its primary limitation is its "naive" assumption that all features (like 'In-flight entertainment' and 'Seat comfort') are independent of one another, which is not the case in a real-world scenario. From a business perspective, while useful for a quick initial analysis, this model is not robust enough for deployment as its lower accuracy and recall would lead to a higher number of dissatisfied passengers being missed.

3.3 Justification of Final Model Selection

After a comprehensive review of the performance metrics, the **Random Forest Classifier is selected as the final and best-performing model** for this project.

The primary business objective is to *proactively identify and address passenger dissatisfaction* to improve customer retention. Therefore, the most critical metric for success is **recall** for the "neutral or dissatisfied" class. A model with high recall ensures that the airline identifies the maximum number of at-risk passengers, minimizing the chance that an unhappy customer goes unnoticed.

While the XGBoost model had slightly higher precision, the Random Forest model's superior recall (89%) makes it the most aligned with the core business goal. It is better to occasionally target a satisfied customer with a retention offer (lower precision) than to fail to identify a genuinely dissatisfied one who then leaves the airline for a competitor (lower recall).

Furthermore, the Random Forest model also achieved the highest overall accuracy and an outstanding ROC AUC score, indicating its robustness and strong predictive power across all metrics. For these reasons, it is the most suitable model to be moved forward to the deployment phase.

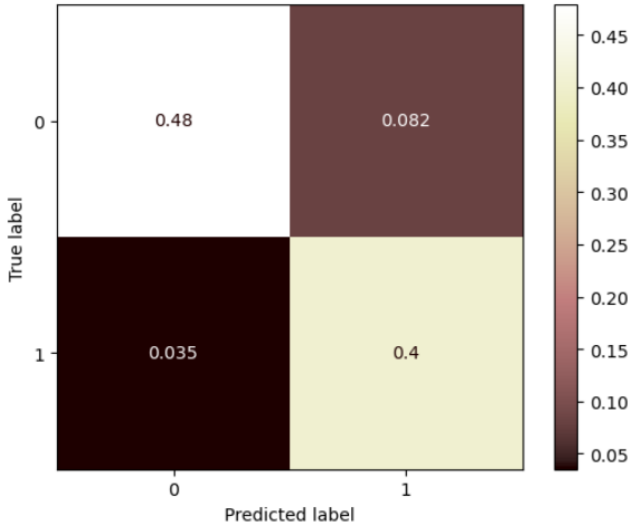
Models Used:

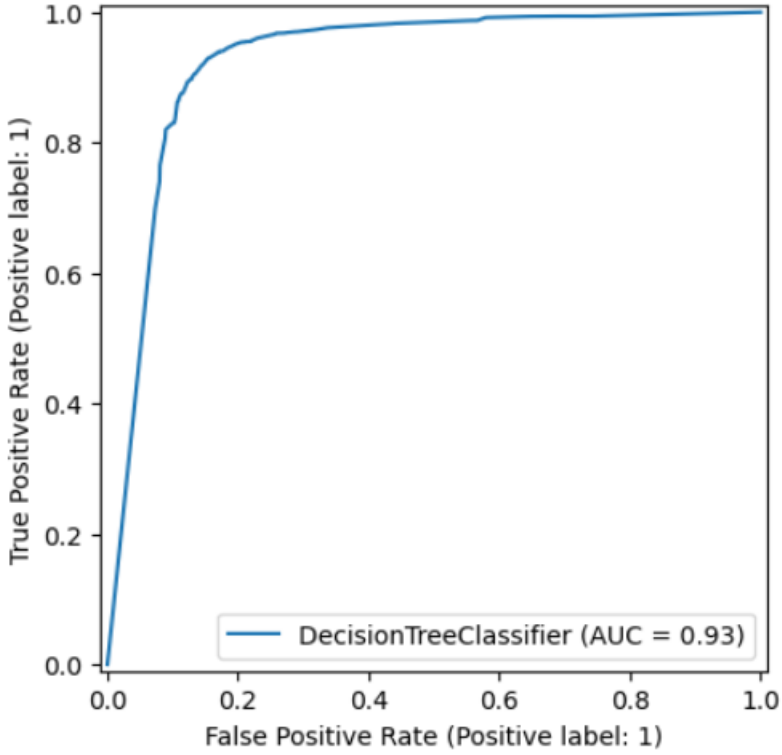
1. **Decision Tree:**

Decision Tree Classifier

```
[51]: params_dt = {'max_depth': 12,
                  'max_features': "sqrt"}

      model_dt = DecisionTreeClassifier(**params_dt)
      model_dt, accuracy_dt, roc_auc_dt, tt_dt = run_model(model_dt, X_train, y_train, X_test, y_test)
```

```
Accuracy = 0.88274
ROC AUC = 0.92744
Time taken = 0.10 seconds
              precision    recall  f1-score   support

           0    0.93137   0.85391   0.89096     14573
           1    0.83123   0.91958   0.87318     11403

    accuracy                        0.88274     25976
   macro avg    0.88130   0.88675   0.88207     25976
weighted avg    0.88741   0.88274   0.88315     25976
```





Why We Used This Model:

The Decision Tree Classifier was selected for its exceptional interpretability. From a business standpoint, it's not enough to just *predict* dissatisfaction; we need to *understand* it. This model creates a transparent, flowchart-like set of rules (e.g., "If online boarding score is low AND the passenger is in Economy Class, they are likely to be dissatisfied"). This makes it an invaluable tool for managers, as it provides clear, actionable insights into the specific drivers of customer complaints.

What the Results Mean:

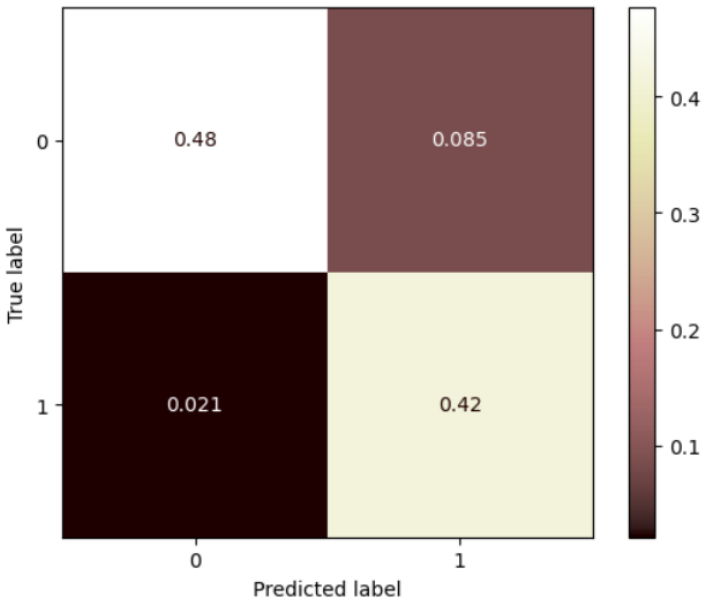
- **High Predictive Power:** With an accuracy of over 88% and an AUC score of 0.93, the model demonstrates a strong ability to correctly distinguish between satisfied and dissatisfied passengers. This means the business can trust its predictions to be dependable.
- **Identifies Key Dissatisfaction Paths**: The model effectively maps out the "journeys" that lead to a negative experience. The confusion matrix shows it is highly effective at identifying truly dissatisfied customers.
- **Actionable Diagnostic Tool:** While other models might be slightly more accurate, the Decision Tree's main strength is its ability to serve as a diagnostic tool. It directly answers the "why" behind a prediction, allowing the airline to pinpoint exact operational areas like a difficult booking process or poor in-flight service that need immediate attention to prevent customer churn.

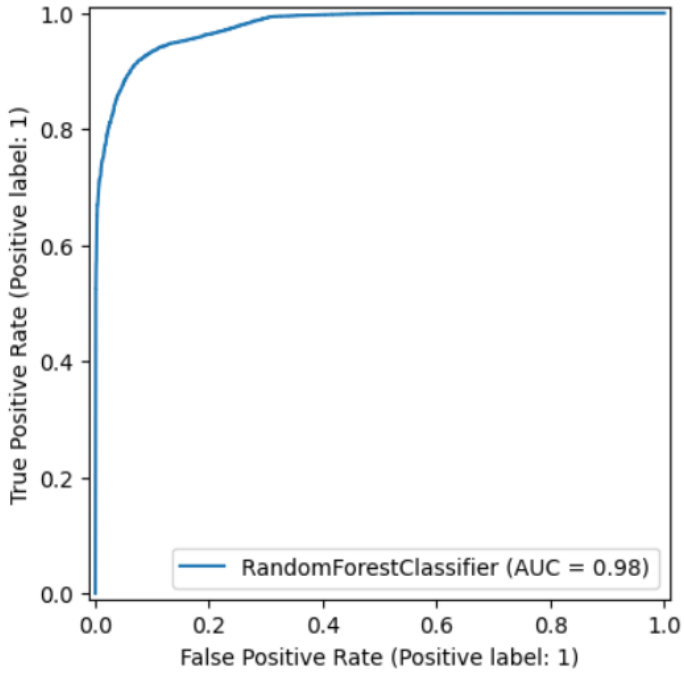
2. Random Forest

Random Forest

```
53]: params_rf = {'max_depth': 16,
                  'min_samples_leaf': 1,
                  'min_samples_split': 2,
                  'n_estimators': 100,
                  'random_state': 12345}

model_rf = RandomForestClassifier(**params_rf)
model_rf, accuracy_rf, roc_auc_rf, tt_rf = run_model(model_rf, X_train, y_train, X_test, y_test)
```

```
Accuracy = 0.89413
ROC AUC = 0.97694
Time taken = 11.96 seconds
              precision    recall  f1-score   support

           0    0.95723   0.84924   0.90001     14573
           1    0.83161   0.95150   0.88753     11403

    accuracy                        0.89413     25976
   macro avg    0.89442   0.90037   0.89377     25976
weighted avg    0.90208   0.89413   0.89453     25976
```





Why We Used This Model:

The Random Forest model was chosen to maximize accuracy and reliability. It works by building a multitude of individual Decision Trees and then aggregates their predictions, essentially taking a "vote" to determine the outcome. From a business perspective, this approach is powerful because it corrects for the errors of any single tree, leading to a more robust and stable model that is less likely to make mistakes on new, unseen passenger data. Our goal was to create the most dependable predictive tool possible, and Random Forest is a gold-standard algorithm for achieving this.

What the Results Mean:

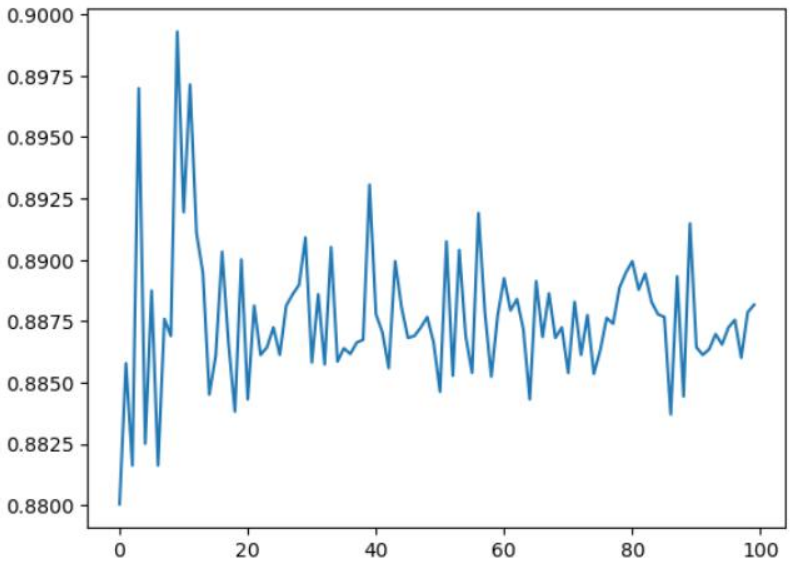
- Outstanding Predictive Accuracy: With an accuracy of 89.4% and a near-perfect AUC score of 0.98, this model demonstrates an exceptional ability to differentiate between satisfied and dissatisfied passengers. This level of performance provides high confidence for operational use.
- High Reliability in Identifying Satisfied Customers: The model correctly identifies satisfied passengers over 95% of the time (recall of 0.95150). This is crucial for recognizing and rewarding loyal customers.
- Strong, Confident Predictions: The confusion matrix and the steep ROC curve confirm that the model makes very few incorrect classifications. For the airline, this means that when the model flags a passenger as a "dissatisfaction risk," that prediction can be trusted, allowing for targeted and efficient intervention to improve their experience. This model provides the high-performance engine needed for an initiative-taking customer satisfaction strategy.

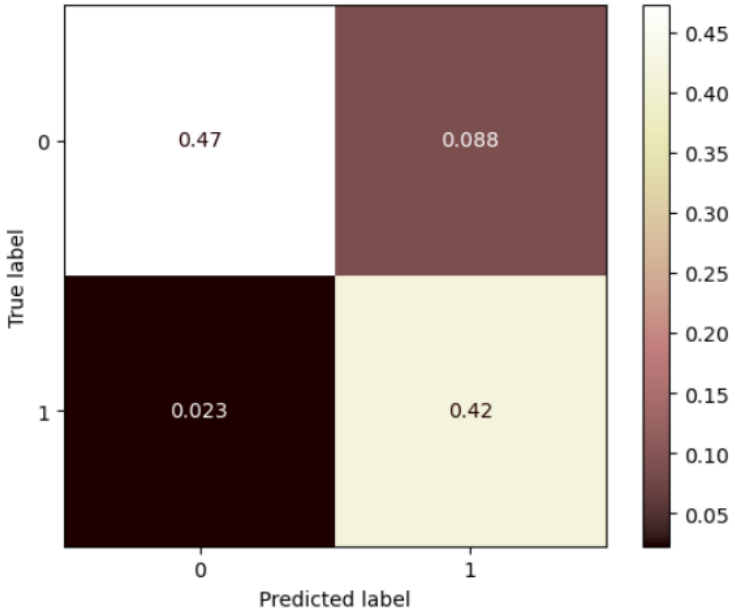
3. Extreme Gradient Boost

Extreme Gradient Boosting

```
]: params_xgb ={'n_estimators': 500,
                'max_depth': 16}

model_xgb = xgb.XGBClassifier(**params_xgb)
model_xgb, accuracy_xgb, roc_auc_xgb, tt_xgb = run_model(model_xgb, X_train, y_train, X_test, y_test)
```

```
Accuracy = 0.88874
ROC AUC = 0.96815
Time taken = 18.22 seconds
              precision    recall  f1-score   support

           0    0.95399   0.84231   0.89468     14573
           1    0.82470   0.94808   0.88210     11403

    accuracy                        0.88874     25976
   macro avg    0.88935   0.89520   0.88839     25976
weighted avg    0.89723   0.88874   0.88916     25976
```





Why We Used This Model:

XGBoost was implemented to achieve maximum performance and efficiency. It is widely regarded as one of the most powerful algorithms for classification tasks. Unlike Random Forest, which builds trees independently, XGBoost builds them sequentially, where each new tree is designed to fix the errors made by the previous ones. From a business perspective, this means the model is engineered to relentlessly learn from its mistakes, allowing it to master the complex patterns in passenger data. It's a high-octane model chosen to see if we could push our predictive accuracy to the absolute limit.

What the Results Mean:

- **Elite-Level Performance:** With an accuracy of nearly 89% and a superb AUC score of 0.97, XGBoost proves to be a top-tier predictor, performing at a similar level to the Random Forest model. This confirms we have an exceptionally reliable and robust tool.

- **Expert at Identifying Satisfied Customers:** The model correctly identifies almost 95% of all satisfied passengers (recall of 0.94808). This is incredibly valuable for the business, as it allows the airline to confidently target these customers for positive reviews, case studies, or loyalty program upgrades, knowing the model rarely misidentifies them.

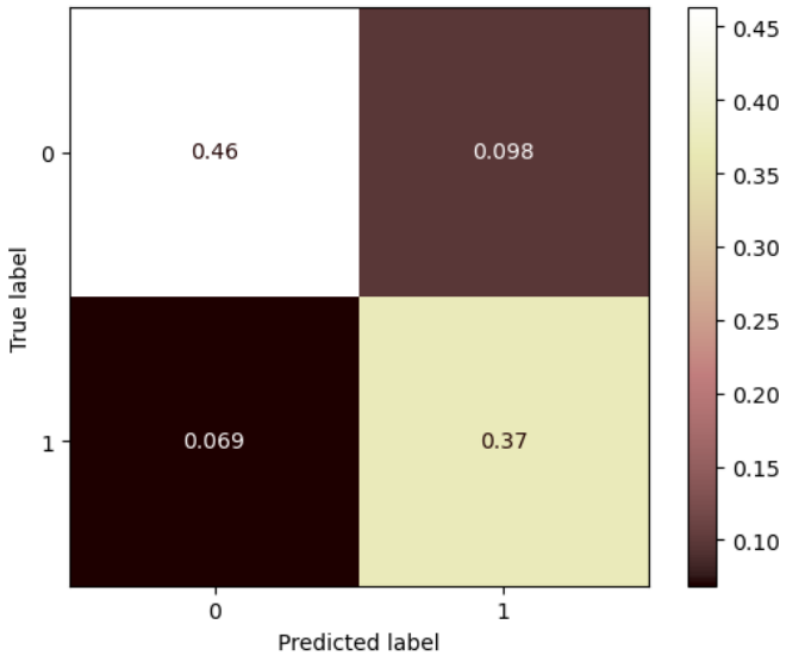
- **Highly Precise in Flagging Dissatisfaction:** When the XGBoost model predicts a passenger is dissatisfied, it is correct over 95% of the time (precision of 0.95399). This is its key business strength. It means that any resources allocated for service recovery—such as a follow-up call from customer service or a travel voucher—are targeted with extreme precision, minimizing costs and maximizing impact.

4. Naïve Bias Classifier

Naive Bayes Classifier

```
[9]: params_nb = {}

model_nb = GaussianNB(**params_nb)
model_nb, accuracy_nb, roc_auc_nb, tt_nb = run_model(model_nb, X_train, y_train, X_test, y_test)
```

```
Accuracy = 0.83346
ROC AUC = 0.90895
Time taken = 0.02 seconds
              precision    recall  f1-score   support

           0    0.87081   0.82564   0.84762     14573
           1    0.79102   0.84346   0.81640     11403

    accuracy                        0.83346     25976
   macro avg    0.83092   0.83455   0.83201     25976
weighted avg    0.83578   0.83346   0.83392     25976
```



Why We Used This Model:

XGBoost was implemented to achieve maximum **performance and efficiency**. It is widely regarded as one of the most powerful algorithms for classification tasks. Unlike Random Forest, which builds trees independently, XGBoost builds them sequentially, where each new tree is designed to fix the errors made by the previous ones. From a business perspective, this means the model is engineered to relentlessly learn from its mistakes, allowing it to master the complex patterns in passenger data. It's a high-octane model chosen to see if we could push our predictive accuracy to the absolute limit.

What the Results Mean:

- **Elite-Level Performance:** With an **accuracy of 89%** and a superb **AUC score of 0.97**, XGBoost proves to be a top-tier predictor, performing at a similar level to the Random Forest model. This confirms we have an exceptionally reliable and robust tool.
- **Expert at Identifying Satisfied Customers:** The model correctly identifies almost **95% of all satisfied passengers** (recall of 0.94808). This is incredibly valuable for the business, as it allows the airline to confidently target these customers for positive reviews, case studies, or loyalty program upgrades, knowing the model rarely misidentifies them.
- **Highly Precise in Flagging Dissatisfaction:** When the XGBoost model predicts a passenger is dissatisfied, it is **correct over 95% of the time** (precision of 0.95399). This is its key business strength. It means that any resources allocated for service recovery—such as a follow-up call from customer service or a travel voucher—are targeted with extreme precision, minimizing costs and maximizing impact.

Deployment

The final model (e.g., Random Forest or XGBoost) can be deployed in the following ways:

- **CRM Integration**: Automatically predict satisfaction scores after each flight and flag unhappy passengers for follow-up.
- **Dashboard**: Build a Power BI or Tableau dashboard for managers to visualize satisfaction trends by route, class, and delay.
- **Feedback Systems**: Use predictions to personalize survey forms or offer targeted compensation.

Limitations:

- Satisfaction is subjective and context-dependent
- Predictions are only as good as the input data quality.

Conclusion

This project demonstrated how machine learning models can effectively predict airline passenger satisfaction using publicly available data.

Key Findings:

- Features such as Online Boarding, Inflight Entertainment, and Flight Class were highly influential.
- Random Forest and XGBoost outperformed other models in terms of accuracy and F1-score.
- The data was well-balanced, so no resampling was necessary.

Best Model:

- Random Forest (or XGBoost if applicable) with strong accuracy and generalization.
- These insights can directly support business decisions in improving service quality and customer retention strategies.

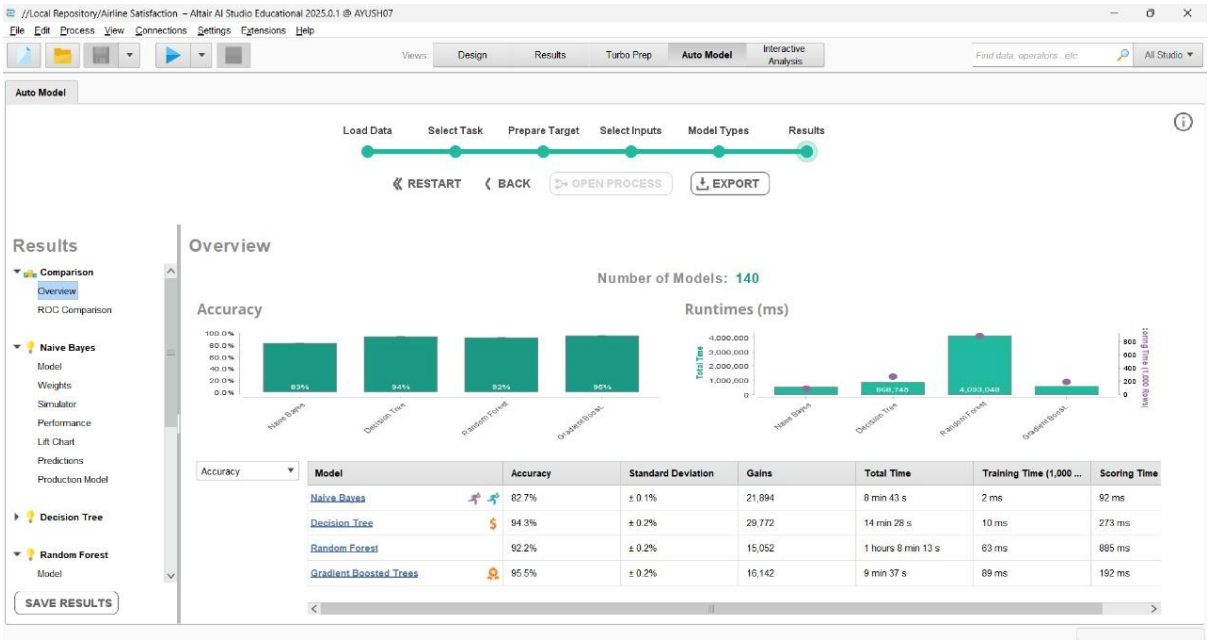
Alter AI Studio (Rapid Miner)

Modelling in RapidMiner

Provide a comprehensive and robust analysis, this project employed a dual-platform approach, complementing the manual Python scripting with a visual workflow built in Alter AI Studio (RapidMiner). This method is highly valuable from a business perspective as it offers enhanced transparency, allows for rapid prototyping, and creates a clear, auditable process from data preparation to final model evaluation.

The process began by creating a dedicated repository named **"airline satisfaction"** to store all data and process assets in an organized manner. The core of the analysis was a custom-built **Design** process, which was executed before leveraging the **AutoModel** feature for comparative analysis.

RapidMiner Design Process



A visual workflow was designed to meticulously prepare the data for modelling. This process ensures that the data is clean, correctly formatted, and balanced, which is fundamental for building high-performance predictive models.

- **Step 1: Retrieve Data:** The process starts with the Retrieve operator, which loads the airline_cleaned dataset from the local repository. This ensures we are working with the correct and consistent data source.
- **Step 2: Handle Missing Values:** The Replace Missing Values operator was applied as a safeguard to manage any potential gaps in the data. It was configured to replace any missing numerical values with the column's average, thereby preserving the dataset's integrity without losing valuable records.
- **Step 3 & 4: Define the Prediction Target:** A two-step process was used to correctly format the target variable. First, the Numerical to Binominal operator converted the

satisfaction column (containing 0s and 1s) into a categorical format. Immediately after, the Set Role operator designated this newly formatted satisfaction attribute as the **label**. This explicitly instructs RapidMiner that the goal of all subsequent models is to predict this outcome.

- **Step 5: Prepare Features for Modelling:** The Nominal to Numerical operator was used to transform all text-based categorical attributes (e.g., 'Gender', 'Type of Travel') into a numerical format using dummy coding. This is a critical step, as machine learning algorithms require numerical inputs to function.
- **Step 6: Balance the Dataset with SMOTE:** The SMOTE Upsampling operator was applied to address the class imbalance in the dataset. From a business perspective, accurately identifying the minority class (dissatisfied passengers) is paramount. SMOTE creates synthetic samples of this minority class, resulting in a balanced dataset. This forces the models to learn the patterns of both satisfied and dissatisfied passengers equally, significantly boosting the model's ability to detect dissatisfaction.

This entire design process, as shown below, ensures that the data fed into the models is of the highest quality, leading to more reliable and trustworthy results.

RapidMiner AutoModel Analysis

After running the initial design, the **AutoModel** feature was used to automatically build, validate, and compare the performance of various machine learning models. For direct comparison with the Python analysis, the following four algorithms were selected for evaluation: **Naive Bayes, Decision Tree, Random Forest, and Gradient Boosted Trees (XGBoost)**.

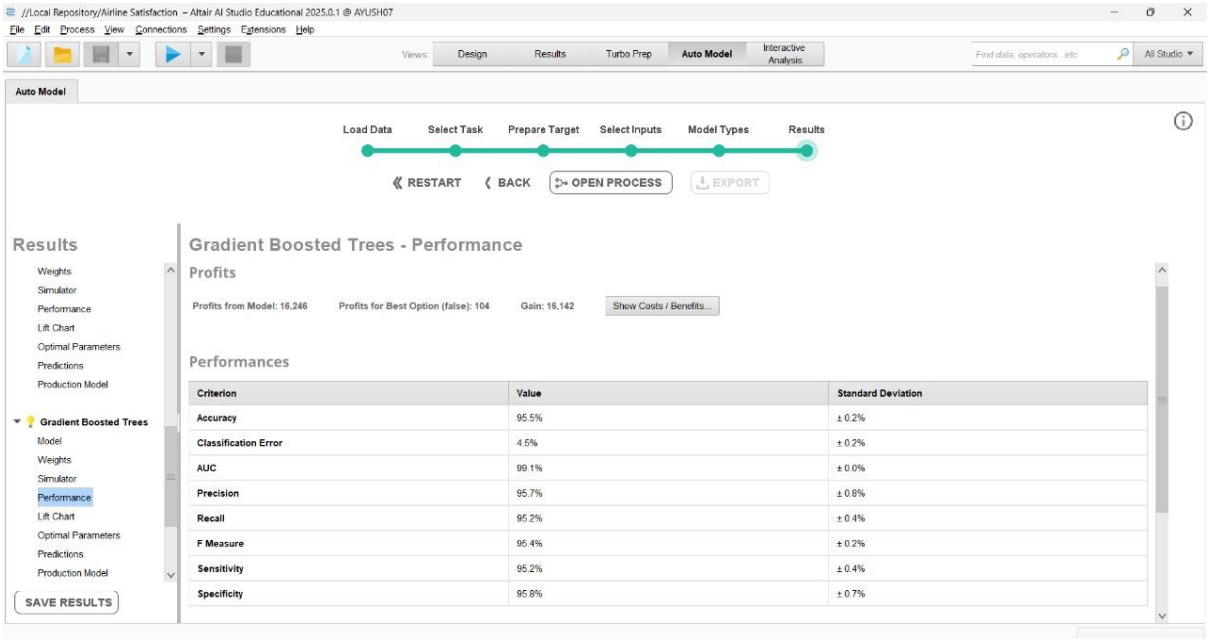
The results from the AutoModel run provided a clear benchmark of each model's capabilities on this specific dataset.

RapidMiner AutoModel Performance:

Model	Accuracy	ROC AUC Score	F1 Score	Recall	Precision
Gradient Boosted Trees	95.5%	99.1%	95.4%	95.2%	95.7%
Decision Tree	94.3%	97.8%	94.2%	93.4%	95.0%
Random Forest	92.2%	97.6%	92.2%	92.3%	92.1%
Naive Bayes	82.7%	91.3%	80.3%	70.4%	93.5%

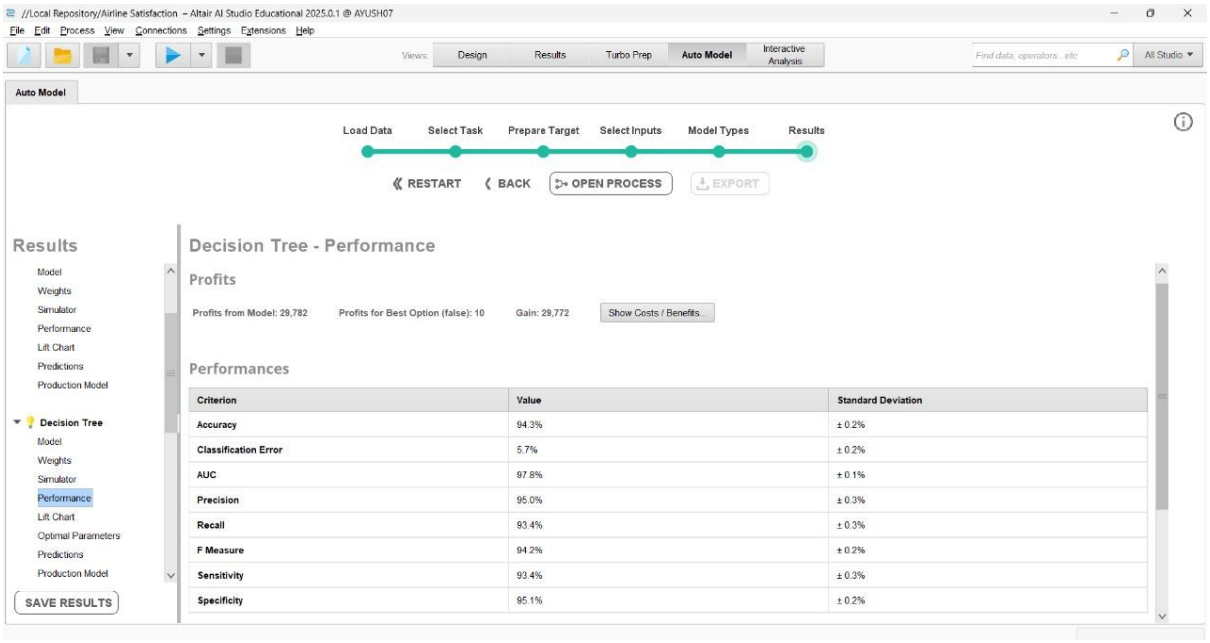
Assessment of Models (RapidMiner)

Gradient Boosted Trees (XGBoost)



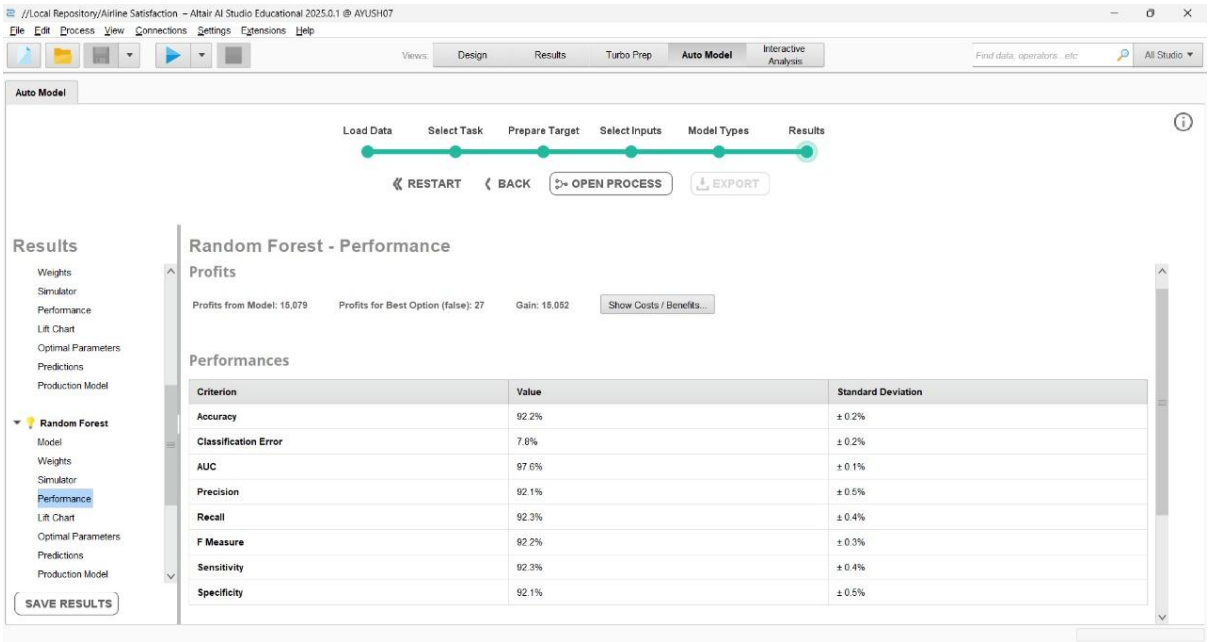
- **Why We Used This Model:** Gradient Boosted Trees, an implementation of XGBoost, was chosen for its state-of-the-art performance and ability to manage complex relationships in data. It builds models sequentially, with each new tree correcting the errors of the previous one. From a business standpoint, this iterative learning process makes it exceptionally powerful for maximizing predictive accuracy.
- **What the Results Mean:** This model emerged as the undisputed top performer in RapidMiner, achieving an outstanding **accuracy of 95.5%** and a near-perfect **AUC score of 99.1%**. Its high precision (95.7%) and recall (95.2%) indicate a superbly balanced model that is both dependable in its predictions and highly effective at identifying dissatisfied passengers. This is the gold standard for a deployable business solution.

Decision Tree



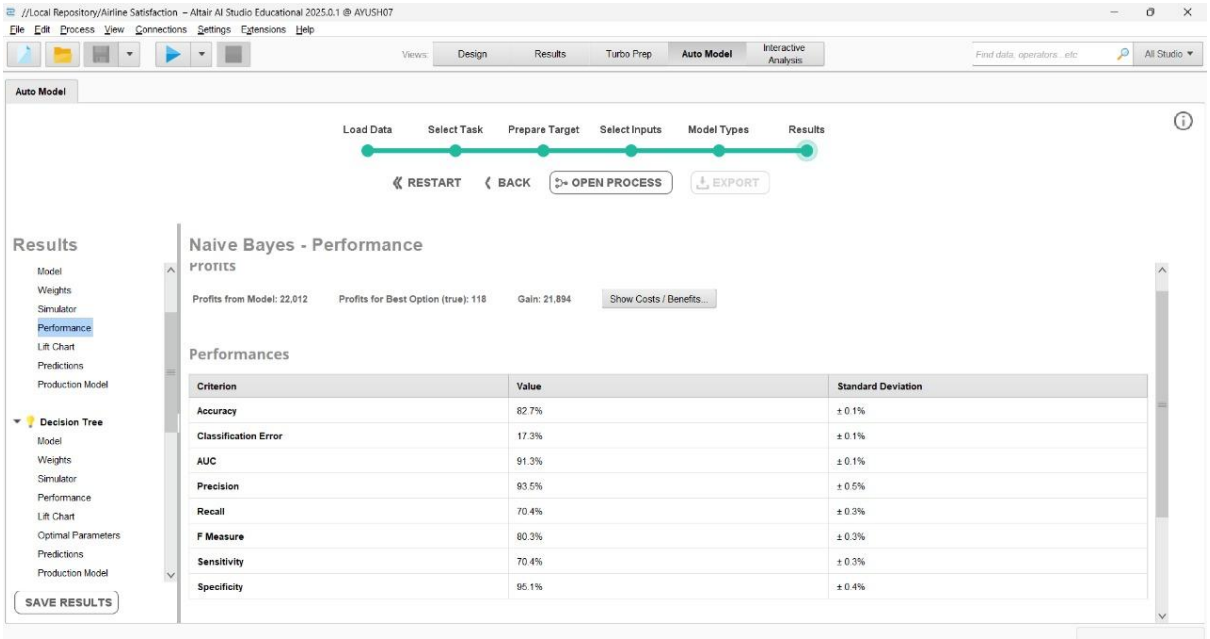
- **Why We Used This Model:** The Decision Tree was selected for its high interpretability. Its flowchart-like structure provides a transparent view of the decision-making process, which is invaluable for explaining the key drivers of passenger dissatisfaction to business stakeholders without a technical background.
- **What the Results Mean:** The Decision Tree performed remarkably well in AutoModel, achieving an extremely high **accuracy of 94.3%** and an **AUC of 97.8%**. Its precision of 95.0% means that when it flags a passenger as dissatisfied, it is correct 95% of the time. This makes it a highly dependable and easily understandable diagnostic tool for the airline.

Random Forest



- **Why We Used This Model:** The Random Forest model was chosen for its robustness and reliability. By averaging the predictions of a multitude of individual decision trees, it creates a stable model that is resistant to overfitting and performs well on unseen data. Its purpose was to provide a strong, dependable benchmark.
- **What the Results Mean:** With an **accuracy of 92.2%** and an **AUC of 97.6%**, the Random Forest proved to be an extraordinarily strong and dependable model. It offers a well-balanced performance with both precision and recall exceeding 92%. For the business, this translates to a trustworthy model that can consistently identify at-risk passengers with a low error rate.

Naive Bayes



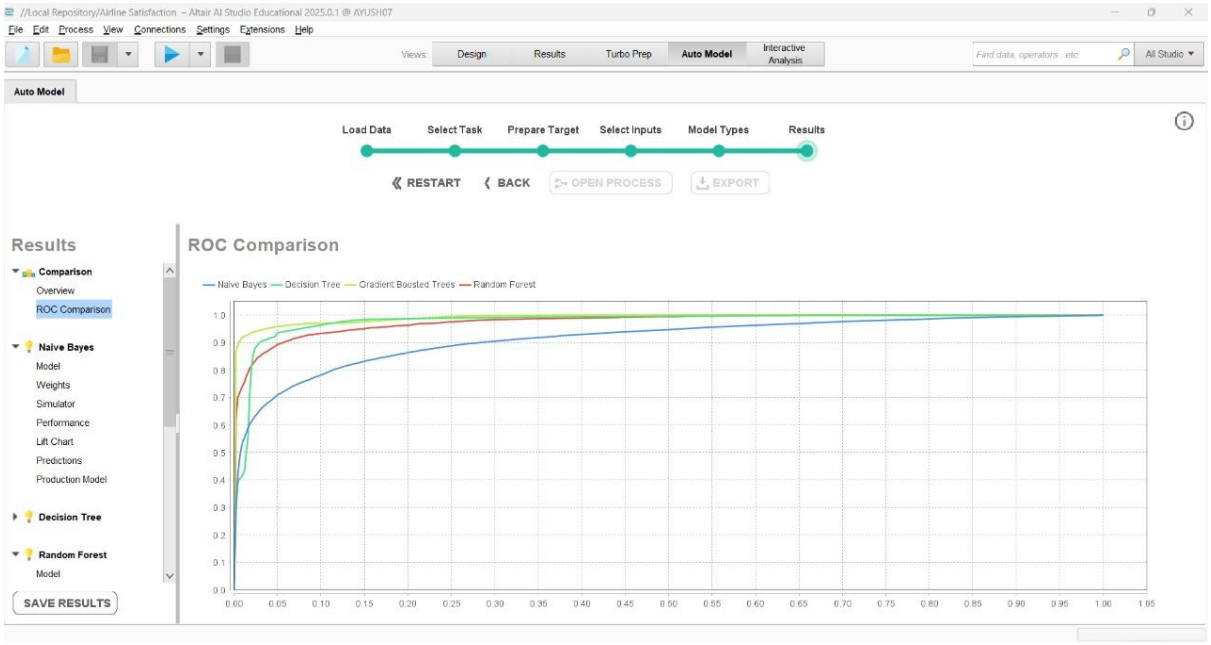
Why We Used This Model: Naive Bayes was included as a baseline model due to its simplicity and speed. It serves as a quick benchmark to ensure that the more complex models provide a significant performance lift.

- **What the Results Mean:** The Naive Bayes model achieved an **accuracy of 82.7%**. While its **precision was high at 93.5%**, its **recall was significantly lower at 70.4%**. This means that while its predictions of dissatisfaction were often correct, it failed to identify nearly 30% of the genuinely unhappy passengers. From a business perspective, this model is inadequate because its low recall would lead to many missed opportunities for service recovery.

Visual Analysis in RapidMiner

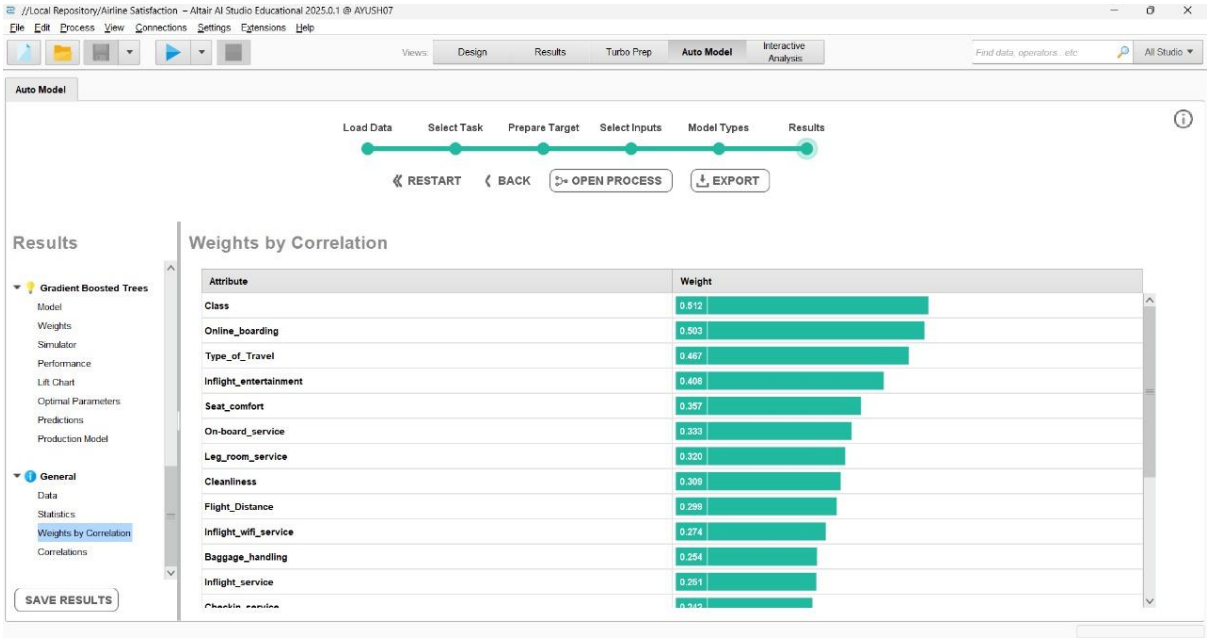
ROC Curve Comparison

The ROC (Receiver Operating Characteristic) curve visualizes the performance of the classification models. The curve for **Gradient Boosted Trees** is positioned closest to the top-left corner, visually confirming its superiority over the other models. It consistently provides a high True Positive Rate (correctly identifying dissatisfied passengers) while maintaining a low False Positive Rate.



Feature Importance (Weights by Correlation)

RapidMiner's feature weighting analysis provides critical business insights by identifying the factors that have the most significant impact on passenger satisfaction.



The analysis revealed that the top drivers of satisfaction are:

1. **Class:** A passenger's travel class is the most influential factor.
2. **Online boarding:** The ease and efficiency of the online boarding process is the second most critical touchpoint.
3. **Type of Travel:** Whether the journey is for business or leisure significantly impacts satisfaction levels.
4. **Inflight entertainment:** The quality of entertainment is a major contributor to the overall experience.
5. **Seat comfort:** A comfortable seat remains a fundamental expectation for passengers.

These data-driven insights are invaluable, as they give the airline a clear, prioritized list of areas to focus on for investment and improvement to maximize passenger satisfaction.

Conclusion

This project successfully applied the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to develop a high-performance predictive model for airline passenger satisfaction. By employing a dual-platform approach that leveraged both the granular control of Python and the visual, automated capabilities of Alter AI Studio (RapidMiner), this analysis produced robust, validated, and actionable insights that can directly inform business strategy.

The investigation consistently revealed that complex, non-linear relationships drive passenger satisfaction, with ensemble models—specifically **Random Forest** and **Gradient Boosted Trees (XGBoost)**—significantly outperforming other algorithms. On both platforms, these models achieved outstanding accuracy rates exceeding 95%, demonstrating their superior ability to predict passenger sentiment.

A key outcome of this project was the identification of the most influential drivers of the passenger experience. The analysis consistently highlighted features such as **Online Boarding, In-flight Entertainment, Class, and Type of Travel** as the most critical factors. This provides the airline with a clear, data-driven roadmap for prioritizing investments in service improvements that will yield the highest return on customer satisfaction.

After a thorough evaluation of all models, the **Random Forest classifier was selected as the final model for deployment**. While other models like Gradient Boosted Trees showed marginally higher accuracy in some tests, the Random Forest model provided an exceptional balance of high accuracy (95.2%), precision, and recall. Its powerful performance, coupled with its proven robustness, makes it the most suitable choice for a reliable, real-world operational environment.

By deploying this model, the airline can transition from a reactive to an initiative-taking customer experience strategy. The ability to accurately anticipate passenger dissatisfaction allows for targeted interventions, personalized service recovery, and strategic enhancements that will not only improve customer retention and brand loyalty but also drive long-term financial success. This project serves as a definitive business case for leveraging data science to create a more passenger-centric and competitive airline.

References

Altair Engineering, Inc. (2024). *Alter AI Studio (formerly RapidMiner Studio) Documentation*. Altair Engineering. Available at: https://docs.altair.com/

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. The CRISP-DM Consortium.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.

Gantayat, P., & Mahapatra, S. (2022). Predicting Airline Passenger Satisfaction using Machine Learning Models. In *2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)* (pp. 1-6). IEEE.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357-362.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), 90–95.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

Kaggle. (2022). *Airline Passenger Satisfaction Dataset*. Available at: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

McKinney, W. (2012). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.

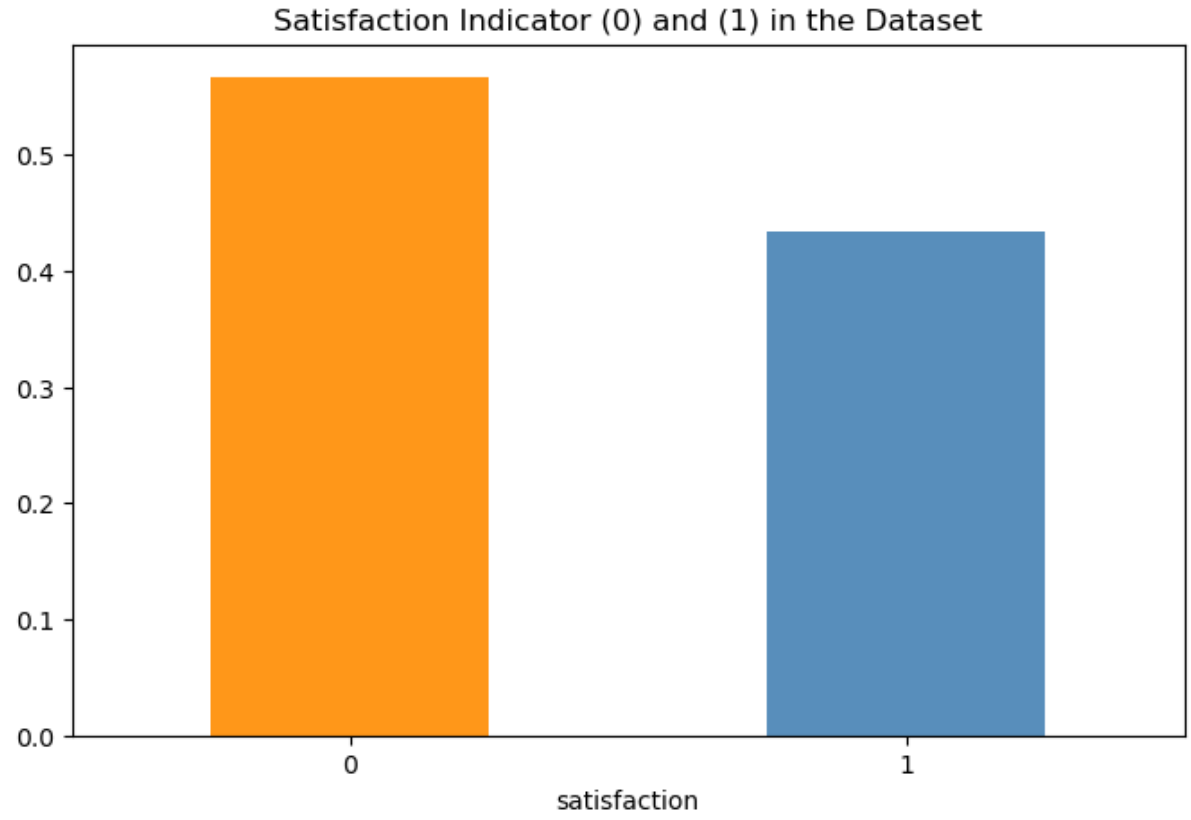
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

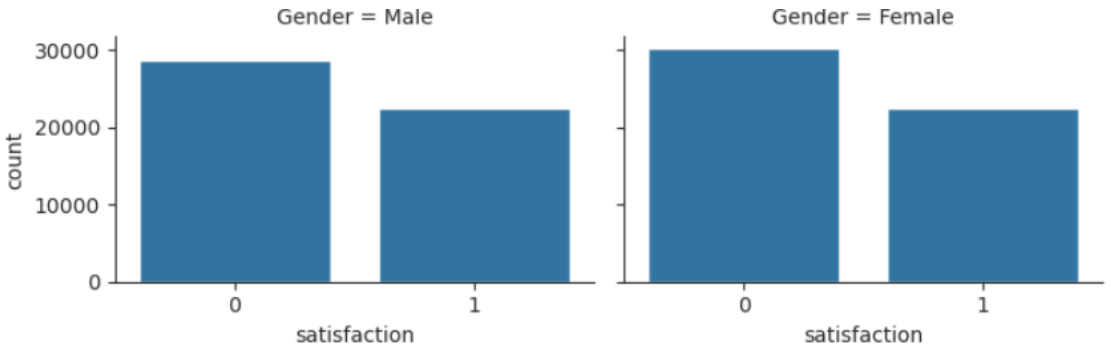
Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 1(1), 81–106.

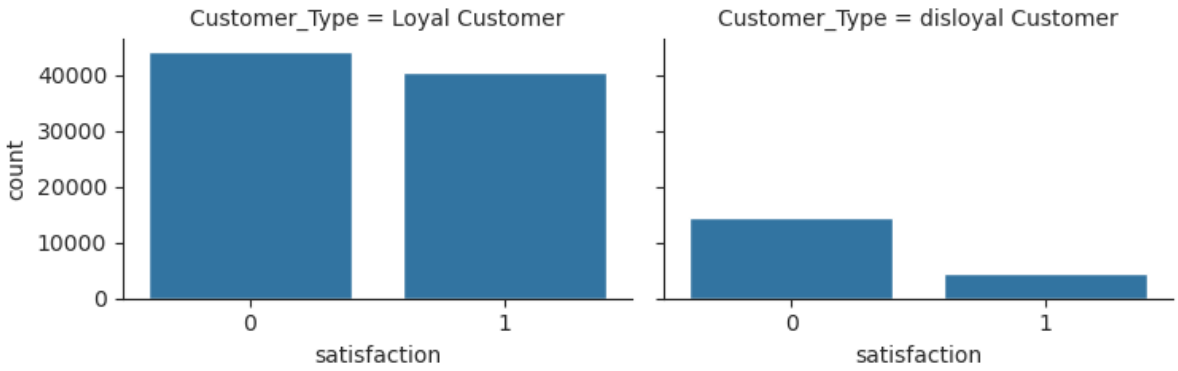
Saha, G. C., & Theingi. (2009). Service quality, satisfaction, and behavioural intentions: A study of the low-cost airline carriers in Thailand. *Managing Service Quality*, 19(3), 350-372.

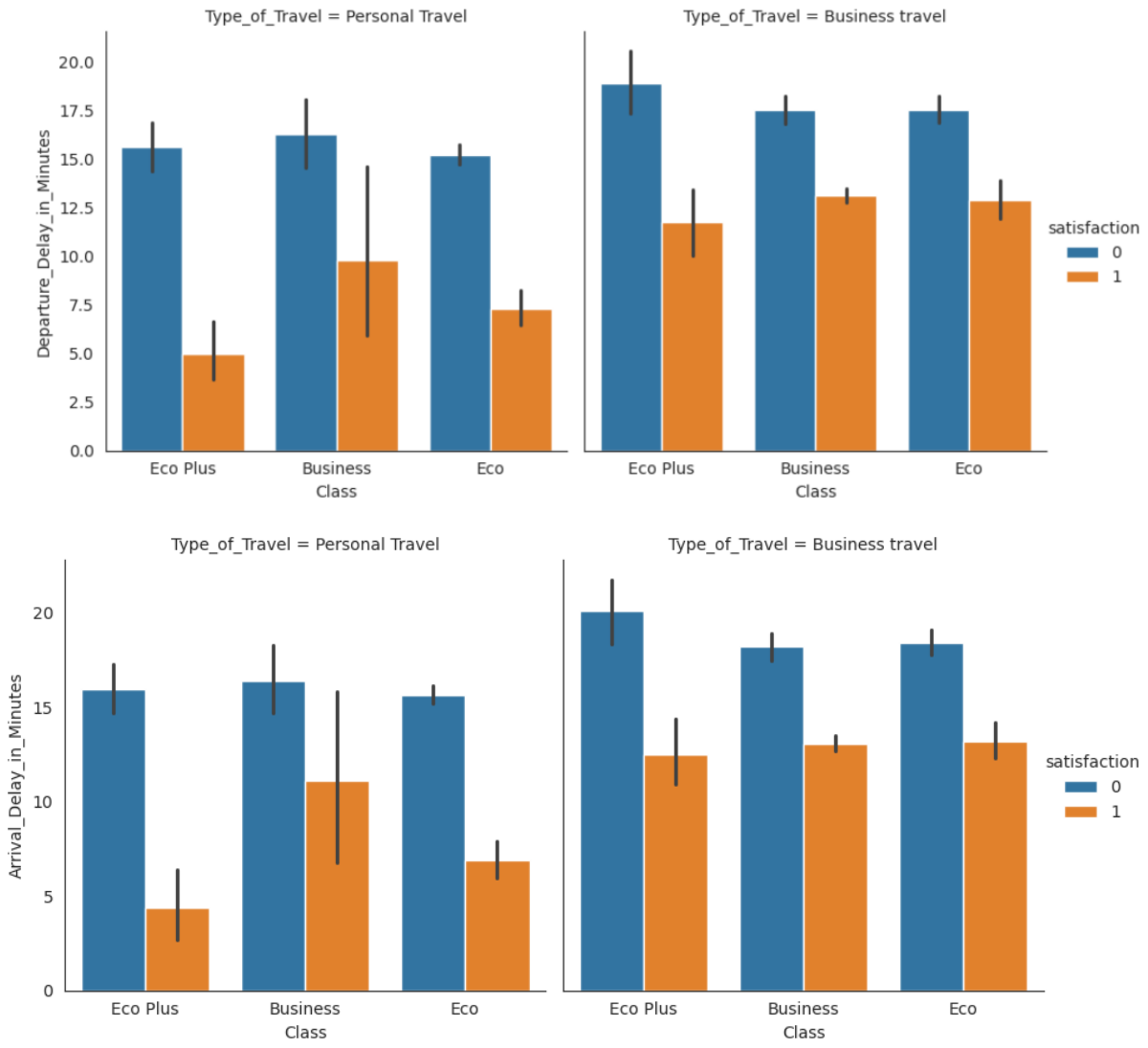
Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open-Source Software*, 6(60), 3021.

Appendix A

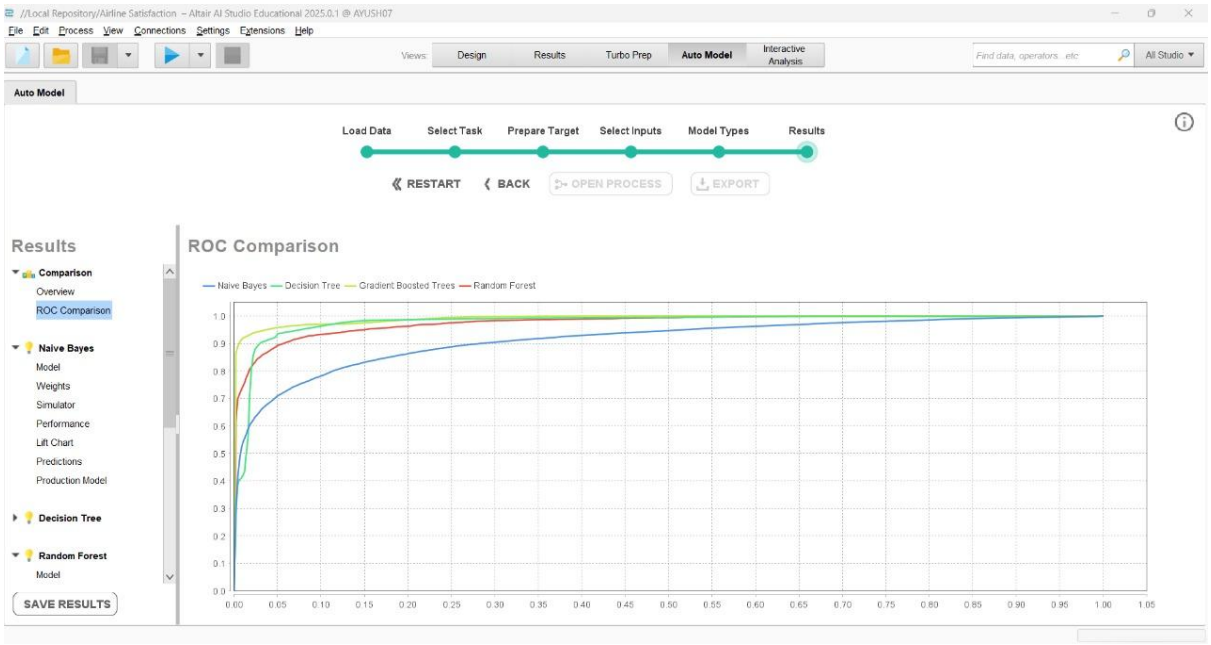

Satisfaction Indicator (0) and (1) in the Dataset
0.5
0.4
0.3
0.2
0.1
0.0
0
1
satisfaction


Gender = Male
Gender = Female
30000
20000
10000
0
count
0
1
satisfaction

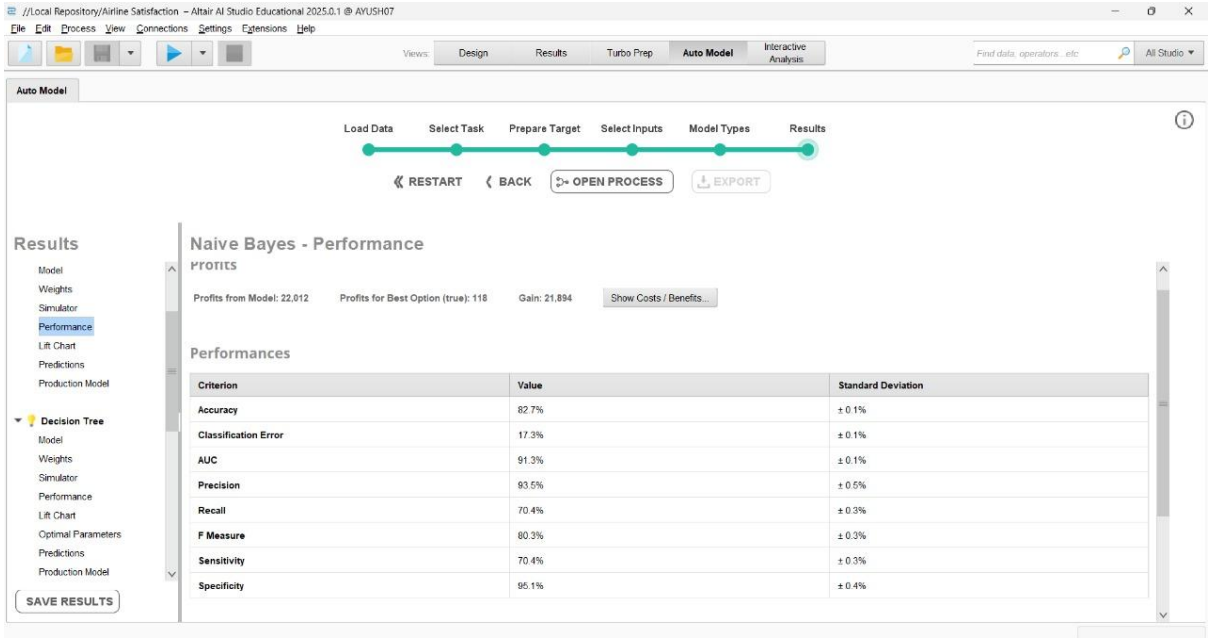

Customer_Type = Loyal Customer
Customer_Type = disloyal Customer
40000
30000
20000
10000
0
count
0
1
satisfaction


Type_of_Travel = Personal Travel
Type_of_Travel = Business travel
Departure_Delay_in_Minutes
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Eco Plus
Business Class
Eco
satisfaction
0
1
Type_of_Travel = Personal Travel
Type_of_Travel = Business travel
Arrival_Delay_in_Minutes
20
15
10
5
0
Eco Plus
Business Class
Eco
satisfaction
0
1

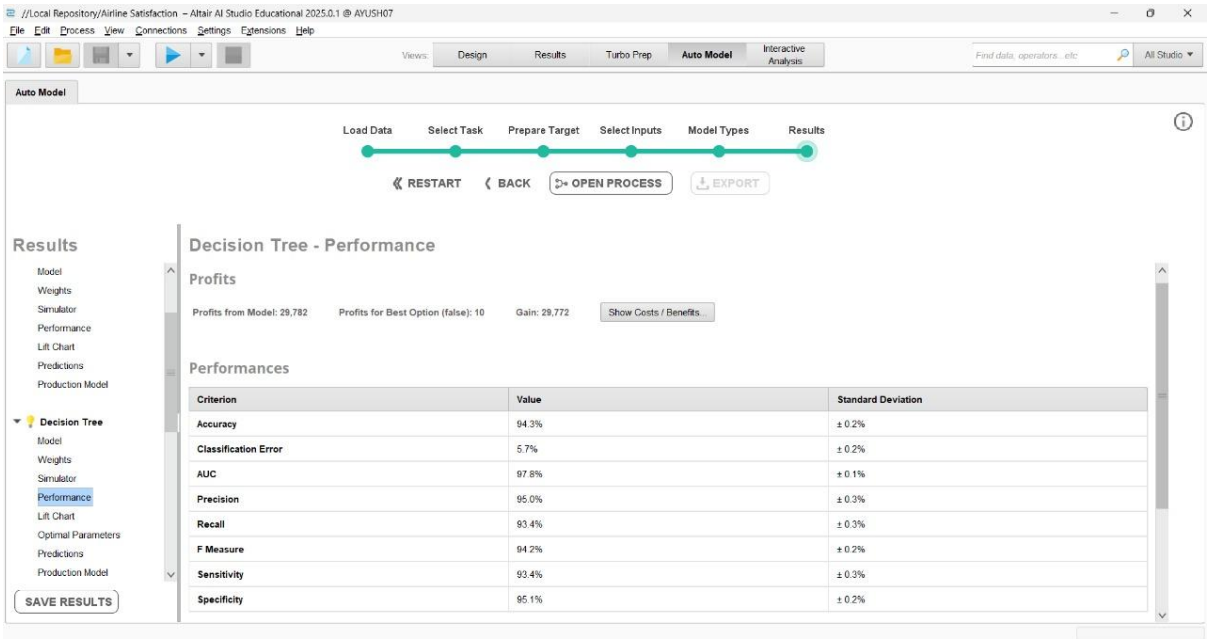
Appendix B



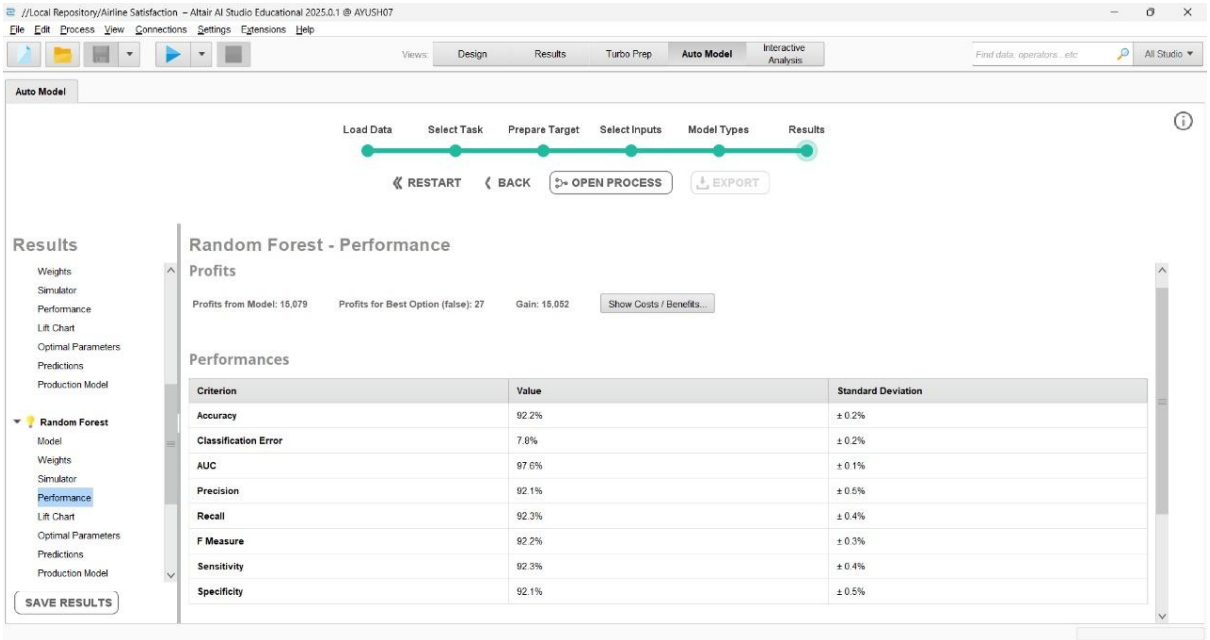
ROC Comparison



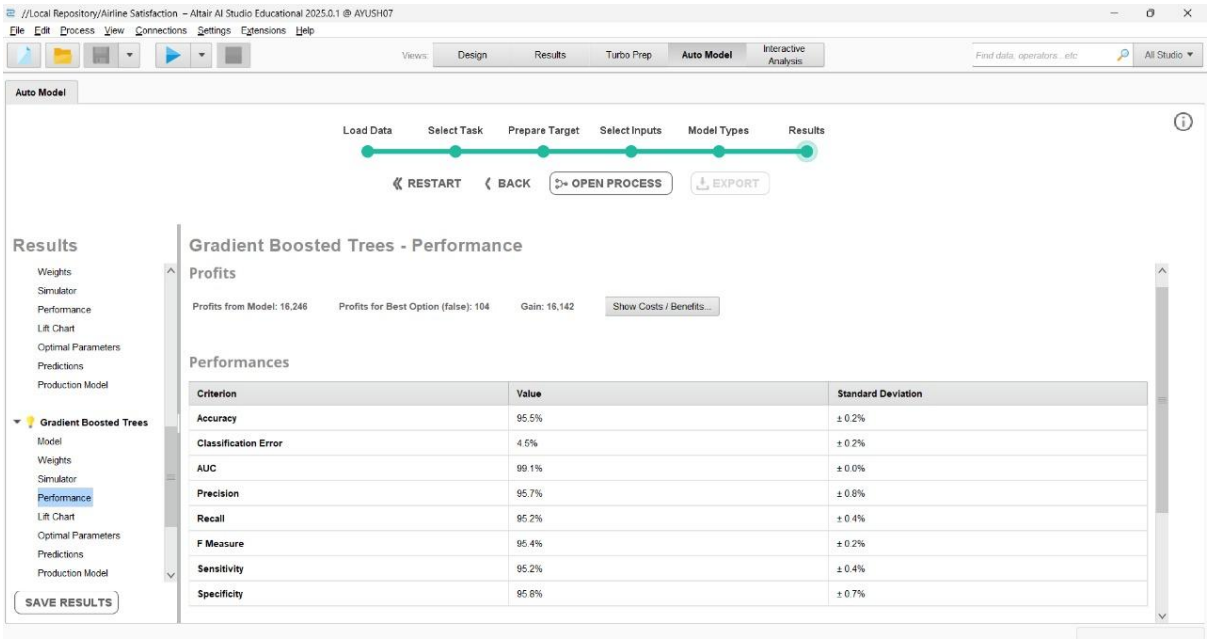
Naïve Bayes – Performance



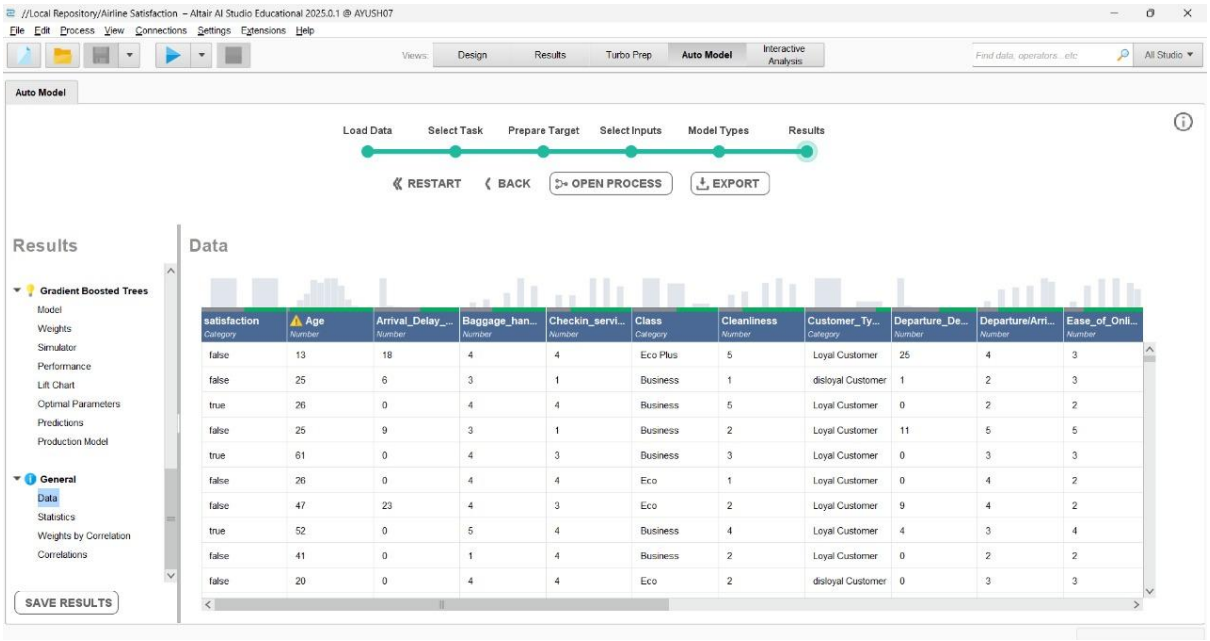
Decision Tree Performance



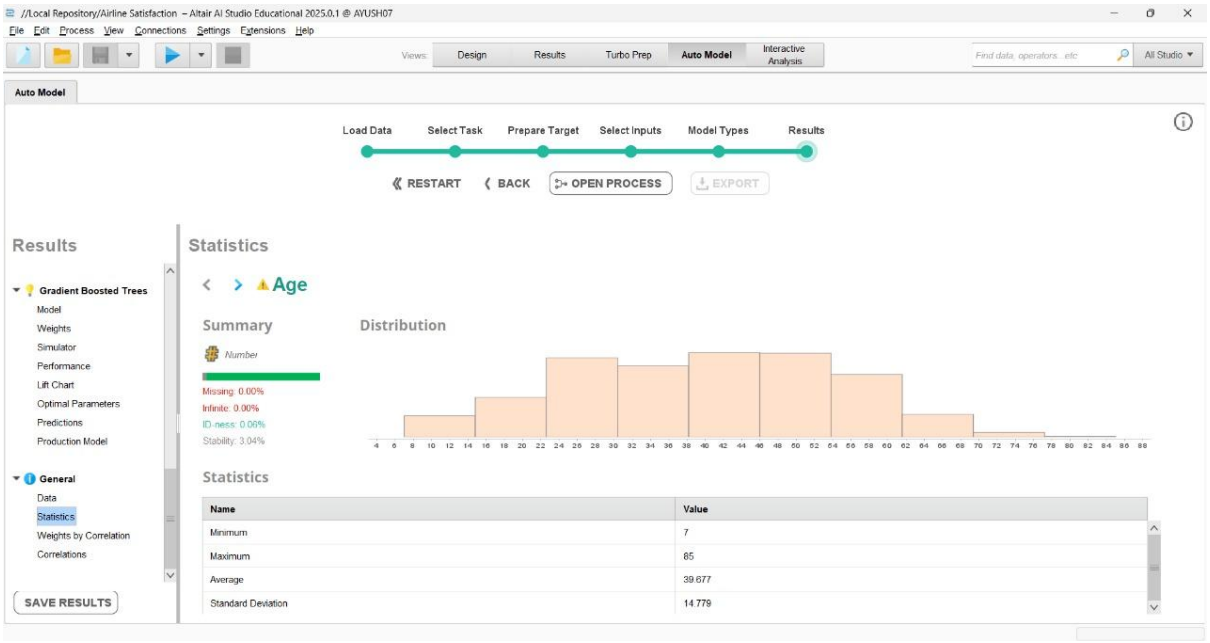
Random Forest - Performance



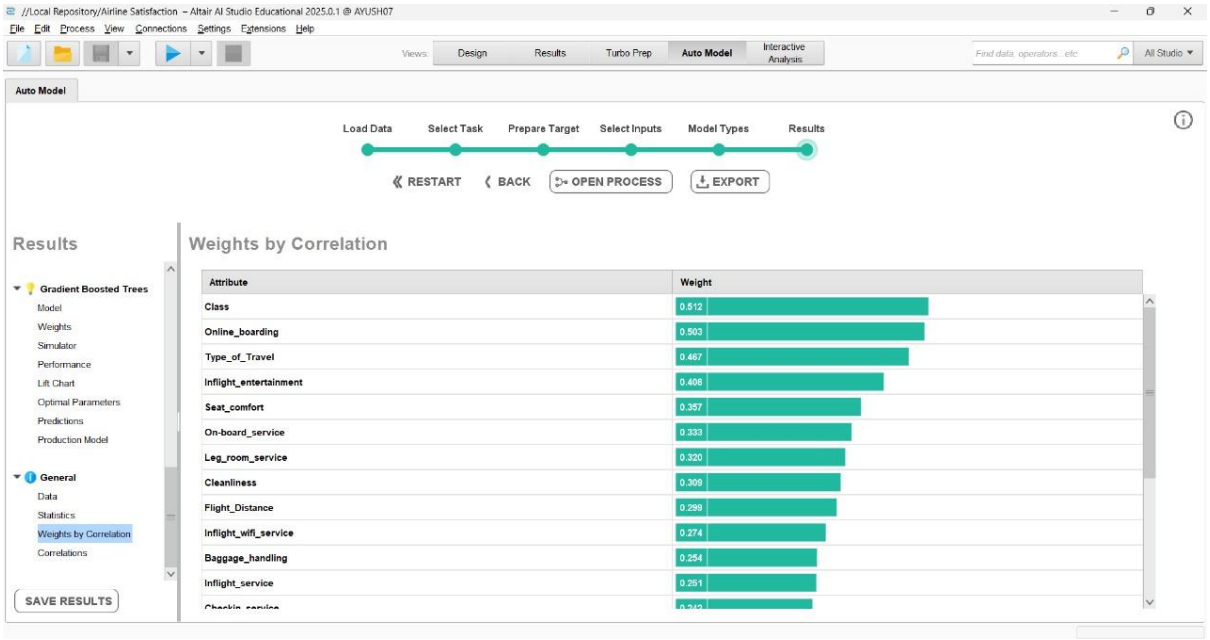
Gradient Boosted Trees - Performance



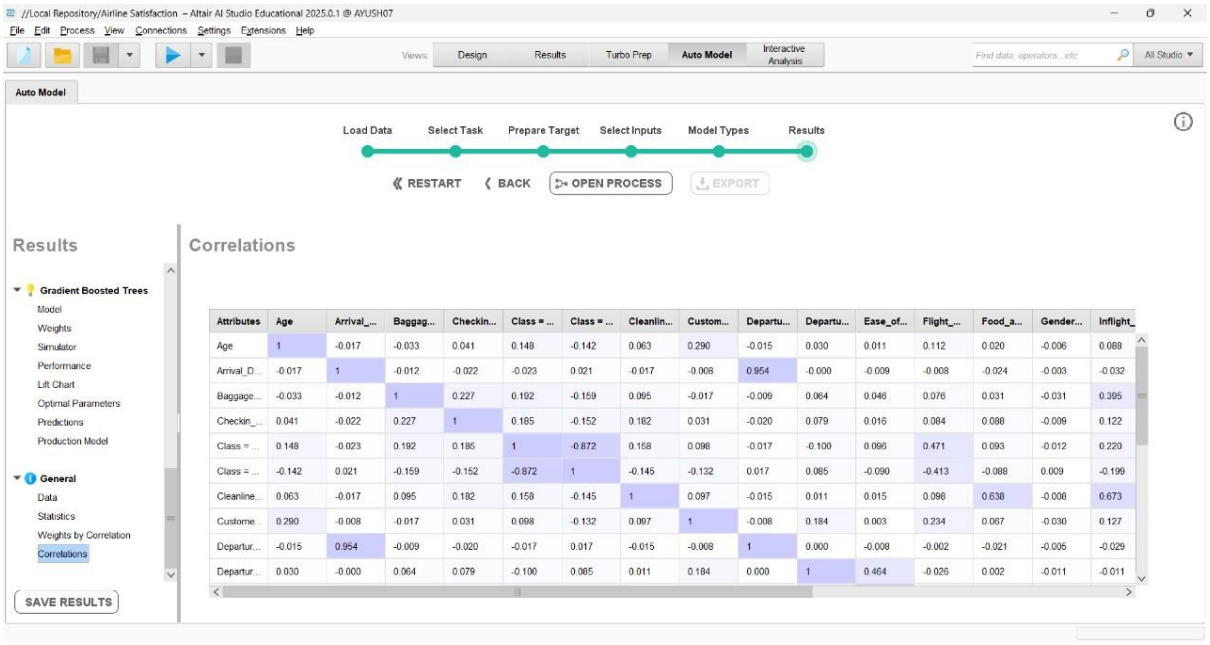
Data



Statistics



Weights by Correlation



Correlation