|  |
| --- |
| BSAN 465 – Predictive Analytics – Course Project |
|  |

[**Pledged**](http://www.stetson.edu/other/honor-system/honorpledge.php)**:** *Alexis Hitchens*

* Project Aligned Learning Outcomes (complete sections based on project):

|  |  |
| --- | --- |
| **BSAN Program Learning Outcomes (PLOs)** | |
| 1. Problem Solving |  |
| 2. Tools and Techniques |  |
| 3. Ethics |  |
| 4. Communication | ✓ |
| **Stetson University** [**General Learning Outcomes**](http://catalog.stetson.edu/undergraduate/general-education/) **(GLOs)** | |
| 1. Writing (WE) | ✓ |
| 2. Information Fluency (WE) |  |
| 3. Speaking (WE) |  |
| 4. Critical Thinking (WE) |  |
| 5. Quantitative Reasoning (Q) |  |
| 6. Knowledge of Human Cultures and the Natural World (A,B,H,S,L,P) |  |
| 7. Personal and Social Responsibility (R,V,W,D,J) |  |
| 8. Integration of Learning |  |

* Employer-Valued Knowledge, Skills, and Abilities (KSAs) Gained:
  + Knowledge…
  + Skills…
  + Abilities…
* Writing Resources
  + [Stetson Writing Resources](http://www.stetson.edu/other/writing-center/resources/writers.php)
  + [Stetson Writing Rubric](https://www.stetson.edu/other/writing-program/media/2017%20Writing%20Outcome%20and%20Rubric%203.0.pdf)

**Course Project Description:**

* Explore an analytics project
* Project and Presentation components include:
* Analytics Document
  + 10 Pages (single-spaced)
  + BEMO/CRISP-DM/SEMMA Methodology
  + Additional resources
    - [Stetson Writing Resources](http://www.stetson.edu/other/writing-center/resources/writers.php)
    - [Stetson Writing Rubric](https://www.stetson.edu/other/writing-program/media/2017%20Writing%20Outcome%20and%20Rubric%203.0.pdf)
* Analytics Presentation
  + 7-minute presentation to the class on the topic explored
  + 1-2 PowerPoint slides each section (title slide + 4 sections)

|  |  |
| --- | --- |
| **Predictive Analytics Case Study** | |
| Title:  (CRISP-DM)  {BEMO}  [SEMMA] | BEMO  **Title:** Customer Satisfactions in the Air  **Dataset:** Invistico\_Airline.csv  **Software:** Analytic Solver for Data Mining |
| Introduction  (Business Understanding)  {Business Opportunity} | *Provide an executive summary*  *-Organizational/Industry background*  *J.D. Power and Associates is a global marketing information services company that conducts surveys of customer satisfaction and product quality. The company was founded by James David Power III in 1968 and has since become a well-known authority in the field of customer satisfaction research.*  *The history of J.D. Power's involvement in airline customer satisfaction can be traced back to the late 1980s. J.D. Power began conducting its annual Airline Satisfaction Study in 1989. This study aimed to measure customer satisfaction with various aspects of the airline industry, including in-flight services, check-in processes, boarding, and baggage handling.*  *Over the years, the methodology and scope of the Airline Satisfaction Study have evolved to keep pace with changes in the airline industry and customer expectations. The study typically involves surveys of thousands of airline passengers, collecting feedback on their recent travel experiences. The results are then analyzed to provide rankings and insights into the performance of different airlines.*  *The airline industry has seen significant changes over the decades, including mergers and acquisitions, advancements in technology, and shifts in consumer preferences. J.D. Power's Airline Satisfaction Study has adapted to these changes, providing valuable benchmarking data for airlines and insights into areas where they can improve the customer experience. (About us history | J.D. power)*  *It's important to note that J.D. Power's studies are just one of many tools used by consumers and industry professionals to assess and compare the performance of airlines.*  *For my project, I wanted to do something similar to the JD Power survey. But first, lets look at the results from 2023. Passenger satisfaction declines overall, mainly due to increased airfare costs, registering a 7-point drop to 791 on a 1,000-point scale, marking the second consecutive year of dissatisfaction. The primary factor is a 17-point decrease in cost and fees from 2022. In contrast, first/business class passengers report a positive trend, with a 9-point increase in satisfaction, attributed to enhanced food and beverage services post-pandemic. Low-cost carriers, particularly in the economy/basic economy segment, face a significant decline of 19 points in satisfaction with costs and fees. An exception to the overall decline is a collective 12-point improvement in food and beverage satisfaction across all segments. JetBlue Airways leads in first/business class satisfaction for the second year (893), followed by Delta Air Lines (865) and United Airlines (848). Delta Air Lines tops premium economy satisfaction (848), with JetBlue Airways (840) and Alaska Airlines (823) following.*  *Southwest Airlines maintains the highest satisfaction in the economy/basic economy segment for the second year (827), with Delta Air Lines (801) and JetBlue Airways (800) in second and third, respectively. The North America Airline Satisfaction Study evaluates eight factors and three segments, collecting responses from 7,774 passengers who flew with major North American airlines in the past month. The study, conducted from March 2022 to March 2023, reveals an overall decrease in passenger satisfaction. (Troy)*  *Business problem/opportunity*   * *Key business objective(s)/goal(s) from predictive modeling*   *My goal is to find which variables from my dataset have the greatest impact on customer satisfaction in the airline industry. With this knowledge, airlines could make improvements in multiple different areas.*  *Customized Services: Utilize predictive analytics to forecast customer preferences and offer personalized services accordingly. For example, if you can anticipate the meal or seating preferences of specific customers, you can provide more tailored options.*  *Operational Improvements: Examine customer satisfaction data to pinpoint pain points in the travel experience, such as delays, lost baggage, or inadequate in-flight entertainment. Addressing these issues can significantly enhance overall satisfaction.*  *Staff Training and Engagement: Analyze data related to customer interactions with airline staff and use this information to customize training programs. Ensure that staff members are well-prepared to address customer needs and concerns effectively.*  *Loyalty Programs: Harness predictive analytics to identify the most effective incentives for customer loyalty. This could involve targeted promotions, personalized offers, or exclusive perks for frequent travelers.*  *Partnerships and Alliances: Collaborate with other businesses to enhance the overall travel experience. This might include forming partnerships with hotels, transportation services, or entertainment providers based on predictive insights into customer preferences.*  *Marketing and Communication: Customize marketing campaigns based on predictive data to emphasize features and services that resonate with the target audience. Effective communication can set realistic expectations and positively influence satisfaction.* |
| Exploration  (Data Understanding /  Data Preparation)  {Exploration}  [Sample, Explore, Modify] | *Provide a results overview following data exploration*   * *Input / Output Variables, Roles, and Types* * *Class:* * *Role: Categorical* * *Type: Nominal* * *Values: Eco = 0, Business = 1, Eco Plus = 2* * *Age:* * *Role: Numeric* * *Type: Continuous* * *Flight Distance:* * *Role: Numeric* * *Type: Continuous* * *Satisfaction Ratings (14 Aspects):* * *Role: Numeric* * *Type: Ordinal* * *Scales: 0-5* * *Seat Comfort* * *Departure/Arrival Time Convenience* * *Food and Drink* * *Gate Location* * *Inflight WiFi Service* * *Inflight Entertainment* * *Online Support* * *Ease of Online Booking* * *On-board Service* * *Leg Room Service* * *Baggage Handling* * *Check-in Service* * *Cleanliness* * *Online Boarding* * *Departure Delay in Minutes:* * *Role: Numeric* * *Type: Continuous* * *Output Variable:* * *Satisfaction:* * *Role: Target/Label* * *Type: Categorical* * *Values: Satisfied = 0, Dissatisfied = 1* * *Data sampling* * *Highlighting the importance of acknowledging manipulated data, I initiated the process by converting all categorical data into a binary set. This transformation extended to four variables, with a particular focus on the key output variable—satisfaction. However, the analytical journey encountered limitations with the use of Analytic Solver, compelling me to navigate through certain constraints in calculations. Challenges arose with variables like customer type, type of travel, and arrival delays, where software-related issues became apparent. Faced with these impediments, a strategic decision was made to selectively eliminate these problematic variables from the dataset under consideration, ensuring a more streamlined and reliable analytical approach.*   *Given the voluminous nature of the original dataset, a sampling strategy was employed, zeroing in on the initial 1800 records for a more manageable and representative subset. This deliberate curation not only facilitated a more nuanced analysis but also helped mitigate the potential impact of outliers and intricacies within the broader dataset. In essence, this meticulous approach aimed at enhancing the overall reliability and effectiveness of the analytical processes.*   * *Graphical data exploration* * *Summary statistics (average, standard deviation, min, max, etc.)* * *“Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.”* *(Patil) Summary statistics play a crucial role in data exploration for predictive analytics. They provide a concise overview of the key characteristics of a dataset, helping analysts and data scientists understand the central tendency, variability, and distribution of the data.* * *A screenshot of a computer    Description automatically generated* * *A table with numbers and text    Description automatically generatedA screenshot of a computer    Description automatically generated* * *Missing data (modify method)* * *Outliers (modify method)*   *In the realm of data analytics, outliers are values in a dataset that significantly deviate from the majority, either being notably larger or smaller. They can signal variations in measurements, experimental errors, or novel occurrences. It is crucial to address outliers in data analytics for two main reasons:*   * *1. Outliers can adversely impact the outcome of an analysis.* * *2. Outliers, or their behavior, may contain valuable information sought by a data analyst.* * *There are two types of outliers:* * *- Univariate outliers are extreme values associated with a single variable.* * *- Multivariate outliers involve unusual or extreme values for at least two variables.* * *Beyond this distinction, outliers fall into different categories:* * *- Global outliers, also known as point outliers, are individual data points significantly distant from the rest of the data distribution.* * *- Contextual outliers, or conditional outliers, are values that deviate significantly within the same context but may not be considered outliers in a different setting, often observed in time series data.* * *- Collective outliers form a subset of data points that differ markedly from the entire dataset.* * *Visualizing data using a box plot simplifies outlier identification. The plot displays a box representing the interquartile range, with outliers shown outside the whiskers. These outliers, situated at the extremes, can be minimum or maximum values. While the instinct may be to eliminate outliers during data cleaning, it's important to note that retaining outliers is sometimes not just beneficial but essential.* * *Removing outliers solely based on their extreme positions can introduce inconsistencies in results, undermining the statistical significance of an analysis. Therefore, careful consideration is needed to ensure that the decision to retain or remove outliers aligns with the analytical goals.* (Sequitin) * *Binary Values and ratings cannot be depicted using a box plot. Therefore, I only had 3 continuous variables. Flight Distance, Departure Delays in Minutes, and Ages. Both Flight Distance and Departure Delay in Minutes both had outliers that lay above the mean. Age did not have any outliers.* * *A graph of flight distance    Description automatically generated* * *A graph with a line graph    Description automatically generated with medium confidence* * *A graph with blue lines    Description automatically generated* * *Data normalization* * *In simple terms, data normalization involves refining collected information to enhance clarity and machine interpretability. Typically, diverse data sources contribute information in various formats, resulting in duplications or irrelevant details, leading to unnecessary storage costs and comprehension challenges.*   *Through the process of data normalization, consistency is established, errors are rectified, and the information is standardized into a uniform format, facilitating easier interpretation and utilization. The primary objective is to minimize redundancy and interdependence within the stored data, ensuring its integrity and eliminating anomalies. (Wrobel)*   * *Relationships between predictors*    + *Correlation Matrix (which variables measure the same thing)*   + *The correlation matrix is a tabular representation illustrating the correlation coefficients between different sets of variables. In statistical terms, correlation coefficients gauge the strength of the relationship between two variables. Each attribute within the dataset is systematically compared with all other attributes, revealing the correlation coefficient for each pair. This analysis is instrumental in identifying pairs with the highest correlation, indicating shared variance within the dataset. Subsequently, a more in-depth examination of these highly correlated pairs can elucidate which attributes are particularly significant for model development. The correlation network depicted displays the relationships among all selected variables, with correlation values ranging from -1 to +1. Variables exhibiting a strong correlation will approach +1, while those with a weaker correlation will tend towards -1. (Dhinakaran)*   *With this correlation matrix, it is color coded, light colors mean less correlation. Dark colors mean more correlated. There is really only one dark color and that is the relationship between satisfaction and departure/arrival time convenience.*  A screenshot of a spreadsheet  Description automatically generated   * *Dimension reduction (PCA)* * *Principal Component Analysis, commonly known as PCA, is a statistical technique designed to condense the information within extensive data tables into a more manageable set of "summary indices." These indices facilitate easier visualization and analysis of the data. Widely recognized as one of the most utilized multivariate statistical methods, PCA finds application in diverse fields such as pattern recognition and signal processing, falling under the broader umbrella of factor analysis. Operating on the principles of projection methods, PCA serves as the foundation for multivariate data analysis.* T*his simplification enables the identification of trends, jumps, clusters, and outliers within the data. PCA accommodates datasets with challenges like multicollinearity, missing values, categorical data, and imprecise measurements. (What is principal component analysis (PCA) and how it is used?)* * *The dataset exhibits notable variability, with the first principal component (PC1) capturing 17.13% of the total variance. Subsequent components contribute to a cumulative variance of 80.86%, indicating a diverse range of information present in the original variables. Certain variables strongly influence specific principal components. For example, "Inflight wifi service" and "Online support" display substantial negative loadings on PC1, emphasizing their significant impact on the observed variance. "Class" exhibits high positive loadings on PC7 and PC8, suggesting a strong association with these components. This may indicate that class-related factors play a substantial role in explaining variability in the dataset. Variables related to in-flight experience, such as "Seat comfort," "Inflight entertainment," and "Food and drink," contribute significantly to PC2 and PC3. This underscores the importance of passenger satisfaction with in-flight services. Departure-related variables, including "Departure/Arrival time convenient" and "Departure Delay in Minutes," contribute to PC4 and PC6. This indicates that aspects related to flight timing and punctuality impact the overall dataset variability. Variables like "Online support" and "Ease of Online booking" strongly affect PC1, suggesting that customer support and online booking experiences are critical factors influencing overall satisfaction. The strong loadings of "Class" on PC7 and PC8 highlight the differentiation of classes in terms of passenger experiences. This may have implications for service quality and satisfaction across different classes. The PCA analysis reveals a multidimensional structure within the dataset, emphasizing the diverse factors influencing passenger satisfaction. The identified principal components provide insights into the underlying patterns and key features that contribute most significantly to the observed variance in the data.*   *A table of numbers on a computer screen  Description automatically generated* |
| Model  (Select PA Techniques, Evaluate)  {Model}  [Model, Assess] | *Partition*   * *Train, # records* * *Validation/Test, # records* * *I did a standard 60/40 split.*   *To evaluate the performance and generalizability of a machine learning algorithm, it is crucial to divide the data into training and test sets. This partitioning enables the simulation of the model's performance on new, unseen data, identification of overfitting, optimization of model parameters, and informed decision-making about its effectiveness before deployment. The validation set, through an impartial evaluation, serves as a critical step in refining the model's hyperparameters, selecting the optimal model, and preventing overfitting. Overfitting occurs when a machine learning model excels on the training data but struggles to generalize to novel, unseen data. This happens when the model learns noise or irrelevant patterns from the training set, leading to subpar performance on the test or validation set. (Acharya)*      *Apply a data mining modelA screenshot of a computer  Description automatically generated*  *A screen shot of a graph  Description automatically generated*  *A screenshot of a computer  Description automatically generatedFirst, I ran a classification tree. For this particular data set, it did not run well at all. This model, just like all of them has its pros and cons. “Advantages: Compared to other algorithms decision trees requires less effort for data preparation during pre-processing. A decision tree does not require normalization of data. A decision tree does not require scaling of data as well. Missing values in the data also do NOT affect the process of building a decision tree to any considerable extent.*  *A Decision tree model is very intuitive and easy to explain to technical teams as well as stakeholders. Disadvantage: A small change in the data can cause a large change in the structure of the decision tree causing instability. For a Decision tree sometimes calculation can go far more complex compared to other algorithms. Decision tree often involves higher time to train the model. Decision tree training is relatively expensive as the complexity and time has taken are more.*  *The Decision Tree algorithm is inadequate for applying regression and predicting continuous values.”* (K) I got horrible numeric results, so I decided to move on to the next model.  The next model I ran was a naïve bayes.   * *A screen shot of a graph    Description automatically generatedA screenshot of a computer    Description automatically generated* * *“The Naïve Bayes classifier is a supervised machine learning algorithm, which is used for classification tasks, like text classification. It is also part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category.” (What are naive Bayes classifiers?) There are some advantages and disadvantages of naïve bayes. “Advantages:* * *Less complex: Compared to other classifiers, Naive Bayes is considered a simpler classifier since the parameters are easier to estimate. As a result, it's one of the first algorithms learned within data science and machine learning courses.* * *Scales well: Compared to logistic regression, Naive Bayes is considered a fast and efficient classifier that is fairly accurate when the conditional independence assumption holds. It also has low storage requirements.* * *Can handle high-dimensional data: Use cases, such document classification, can have a high number of dimensions, which can be difficult for other classifiers to manage.* * *Disadvantages:* * *Subject to Zero frequency: Zero frequency occurs when a categorical variable does not exist within the training set. For example, imagine that we're trying to find the maximum likelihood estimator for the word, "sir" given class "spam", but the word, "sir" doesn't exist in the training data. The probability in this case would zero, and since this classifier multiplies all the conditional probabilities together, this also means that posterior probability will be zero. To avoid this issue, smoothing can be leveraged.* * *Unrealistic core assumption: While the conditional independence assumption overall performs well, the assumption does not always hold, leading to incorrect classifications.” (What are naive Bayes classifiers?)* * *It seems that Class 0 predictions are more prevalent, with 148 instances correctly predicted as 0 and only 1 misclassification as 1. However, there are 374 instances where the model predicted 'nan' when the actual class was 0.* * *For Class 1, there are 60 correct predictions and 137 instances where the model predicted 'nan' instead of 1.* * *There are no instances where 'nan' is correctly predicted, and all predictions for 'nan' are marked as 'nan.' Class 0 has a high error rate of 71.70%, indicating that the model struggles with accurate predictions for this class. Class 1 also has a high error rate of 69.54%. It’s worth noting that the number of errors is substantial relative to the total number of cases, indicating significant room for improvement. The overall accuracy is low at 28.89%, suggesting that the model's predictive performance is not satisfactory. Specificity, Sensitivity, Precision, and F1 score are all marked as #N/A, which indicates that these metrics are not calculable due to the presence of 'nan' values. The warning at the end suggests that there are records in the validation partition that cannot be classified. This could be due to feature values in new records that were not present in the training data. It highlights the importance of ensuring consistency between training and validation data. The model appears to struggle with accurate predictions for both Class 0 and Class 1, and the presence of 'nan' values complicates the evaluation metrics. There is room for improvement in both model training and handling of missing values. I concluded that this was not a good model and moved on.* * *A screenshot of a computer    Description automatically generated* * *A screenshot of a graph    Description automatically generated*   *A screen shot of a graph  Description automatically generatedA screenshot of a computer  Description automatically generated*  *Logistic regression helps us make predictions when the result is binary (yes/no, pass/fail, etc.) by using input features to calculate probabilities.*  *“Advantages:*  *1. Logistic Regression performs well when the dataset is linearly separable.*  *2. Logistic regression is less prone to over-fitting but it can overfit in high dimensional datasets. You should consider Regularization (L1 and L2) techniques to avoid over-fitting in these scenarios.*  *3. Logistic Regression not only gives a measure of how relevant a predictor (coefficient size) is, but also its direction of association (positive or negative).*  *4. Logistic regression is easier to implement, interpret and very efficient to train.*  *Disadvantages of Logistic Regression:*  *1. Main limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. Most of the time data would be a jumbled mess.*  *2. If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise it may lead to overfit.*  *3. Logistic Regression can only be used to predict discrete functions. Therefore, the dependent variable of Logistic Regression is restricted to the discrete number set. This restriction itself is problematic, as it is prohibitive to the prediction of continuous data.” (Jain)*  *True Positives (477): Instances where the model correctly predicted Class 0.*  *False Positives (46): Instances where the model incorrectly predicted Class 1 when the true class was Class 0.*  *False Negatives (43): Instances where the model incorrectly predicted Class 0 when the true class was Class 1.*  *True Negatives (154): Instances where the model correctly predicted Class 1.*  *Class 0 (Success Class):*  *The model performed relatively well, with an 8.80% error rate.*  *Class 1:*  *Class 1 has a higher error rate (21.83%), indicating more difficulty in predicting this class.*  *Accuracy (87.64%): Overall, the model is accurate.*  *Specificity (78.17%): The ability of the model to correctly identify Class 0 instances is good.*  *Sensitivity (Recall - 91.20%): The model performs even better in identifying Class 1 instances.*  *Precision (91.73%): When the model predicts Class 1, it is correct 91.73% of the time.*  *F1 Score (91.47%): A balance between precision and recall, providing a single metric to evaluate the model's overall performance.*  *Intercept (2.67): The log-odds of the response variable being Class 1 when all predictors are zero.*  *Class Coefficient (-0.08): For every one-unit increase in the predictor 'Class,' the log-odds of the response variable being Class 1 decreases by 0.08.*  *Other Coefficients: Similar interpretations can be made for other predictors.*  *Odds Ratios: The odds of the response variable being Class 1 are 14.49 times higher when the intercept is considered.*  *The odds decrease by 8% for every one-unit increase in the 'Class' predictor.*  *The logistic regression model provides coefficients for each predictor variable, which represent the influence of that variable on the log-odds of the target variable (customer satisfaction). The sign of the coefficient indicates the direction of the influence (positive or negative), and the magnitude indicates the strength of the influence.*  *Class:*  *Coefficient: -0.0814*  *Interpretation: As the Class variable increases by 1 unit (moving from Eco to Business, or from Eco Plus to Business), the log-odds of customer satisfaction decrease by 0.0814.*  *Age:*  *Coefficient: -0.4316*  *Interpretation: As Age increases by 1 unit, the log-odds of customer satisfaction decrease by 0.4316.*  *Flight Distance:*  *Coefficient: -0.7826*  *Interpretation: As Flight Distance increases by 1 unit, the log-odds of customer satisfaction decrease by 0.7826.*  *Seat Comfort:*  *Coefficient: 0.3314*  *Interpretation: As Seat Comfort increases by 1 unit, the log-odds of customer satisfaction increase by 0.3314.*  *Departure/Arrival Time Convenience:*  *Coefficient: 3.0387*  *Interpretation: As Departure/Arrival Time Convenience increases by 1 unit, the log-odds of customer satisfaction increase by 3.0387. This has a strong positive influence.*  *Food and Drink:*  *Coefficient: -0.7914*  *Interpretation: As the satisfaction rating for Food and Drink increases by 1 unit, the log-odds of customer satisfaction decrease by 0.7914.*  *Gate Location:*  *Coefficient: 0.3018*  *Interpretation: As the satisfaction rating for Gate Location increases by 1 unit, the log-odds of customer satisfaction increase by 0.3018.*  *Inflight WIFI Service:*  *Coefficient: 0.0597*  *Interpretation: As the satisfaction rating for Inflight WIFI Service increases by 1 unit, the log-odds of customer satisfaction increase by 0.0597.*  *Inflight Entertainment:*  *Coefficient: 0.8679*  *Interpretation: As the satisfaction rating for Inflight Entertainment increases by 1 unit, the log-odds of customer satisfaction increase by 0.8679.*  *Online Support:*  *Coefficient: -0.1019*  *Interpretation: As the satisfaction rating for Online Support increases by 1 unit, the log-odds of customer satisfaction decrease by 0.1019.*  *The variables with positive coefficients (e.g., Departure/Arrival Time Convenience, Inflight Entertainment) have a positive influence on customer satisfaction, while those with negative coefficients (e.g., Age, Flight Distance, Food and Drink) have a negative influence. Seat Comfort and Gate Location also positively influence satisfaction.* |
| Application  (Deploy, Implement)  {Operationalize} | *Provide overall recommendations to business*   * *Model*    + *Model advantages / disadvantages*   + *Performance evaluation*   + *Selection recommendation*   *Model Selection: Linear Regression*  *Advantages:*  *- Well-suited for linearly separable datasets.*  *- Less prone to overfitting.*  *- Provides interpretable coefficients for each predictor.*  *- Efficient to implement and train.*  *Disadvantages:*  *- Assumes linearity between variables.*  *- May overfit in high-dimensional datasets.*  *- \*\*Accuracy: \*\* 87.64%*  *- \*\*Specificity: \*\* 78.17%*  *- \*\*Sensitivity (Recall): \*\* 91.20%*  *- \*\*Precision: \*\* 91.73%*  *- \*\*F1 Score: \*\* 91.47%*  *- \*\*Error Rates: \*\* 8.80% for Class 0, 21.83% for Class 1.*  *Positive coefficients (e.g., Seat Comfort, Inflight Entertainment) positively influence satisfaction.*  *Negative coefficients (e.g., Age, Flight Distance, Food and Drink) negatively influence satisfaction.*  *- Odds of satisfaction are 14.49 times higher when considering the intercept.*  *- Odds decrease by 8% for every one-unit increase in the 'Class' predictor.*  *Logistic Regression is suitable for its interpretability, efficiency, and performance on linearly separable datasets.*  *Be cautious of assumptions of linearity; consider regularization in high-dimensional scenarios.*  *The model demonstrates high accuracy and precision, especially in correctly identifying Class 1 instances (customer satisfaction).*  *- Address the higher error rate in predicting Class 0 instances; explore further optimization opportunities.*  *Focus on aspects with positive coefficients (e.g., Seat Comfort, Inflight Entertainment, Departure/Arrival Time Convenience) to enhance customer satisfaction.*  *Mitigate negative influences (e.g., Age, Flight Distance, Food and Drink) through targeted improvements.*  *- Regularly monitor customer feedback and update the model as needed.*  *- Explore customer sentiment analysis and additional features to enhance predictive capabilities.*  *- Consider refining the model for specific customer segments based on demographics or travel behavior.*  *- Tailor marketing and service strategies to meet the unique needs of different customer groups.*  *- Periodically review the model's performance against evolving business dynamics.*  *- Update predictors or consider more advanced models as the business landscape changes.*  *- Leverage insights from the model to customize staff training programs, focusing on areas identified as crucial for customer satisfaction.*  *- Use predictive insights to form partnerships that align with customer preferences.*  *- Enhance the overall travel experience by collaborating with businesses in related industries.*  *In conclusion, while the logistic regression model provides valuable insights into customer satisfaction, it's essential to continuously refine and enhance strategies based on evolving data and business dynamics. Regular monitoring, targeted improvements, and strategic collaborations can contribute to sustained success in delivering exceptional customer experiences.*   * *Overall summary recommendations*    + *Timeline, resources, phasesat*   + *ROI, implementation cost*   + *Implementation Plan:*   *Timeline:*  *Phase 1 (Months 1-2): Model Implementation and Initial Training*   * + *Implement logistic regression model.*   + *Train the model using historical data.*   *Phase 2 (Months 3-4): Performance Evaluation and Refinement*   * + *Evaluate model performance using real-time data.*   + *Identify areas for improvement and refinement.*   *Phase 3 (Months 5-6): Feature Enhancement and Staff Training*   * + *Enhance model with additional features or customer segments.*   + *Customize staff training based on model insights.*   *Resources:*   * + *Data Scientists/Analysts: Engage professionals to implement, train, and refine the model.*   + *IT Support: Ensure seamless integration of the model into existing systems.*   + *Customer Service Training Team: Collaborate for customizing staff training programs.*   *Phases:*  *Phase 1: Model Implementation*   * + *Develop and implement the logistic regression model.*   + *Train the model using historical customer satisfaction data.*   *Phase 2: Performance Evaluation and Refinement*   * + *Evaluate model performance using real-time data.*   + *Identify and address areas for improvement.*   *Phase 3: Feature Enhancement and Staff Training*   * + *Enhance the model with additional features or customer segmentation.*   + *Customize staff training programs based on model insights.*   *Return on Investment (ROI):*  *Cost Considerations:*   * + *Initial Implementation Cost: Invest in hiring data scientists and IT support for model implementation. Estimate the cost of staff training programs.*   + *Ongoing Monitoring and Refinement: Allocate resources for continuous monitoring, refinement, and updates.*   *Revenue Generation:*   * + *Increased Customer Satisfaction: Expect improved customer satisfaction, leading to increased loyalty and repeat business.*   *Targeted Marketing: Use personalized insights for targeted marketing, potentially increasing customer acquisition.*  *Quantifiable Metrics:*   * + *Trackable Customer Satisfaction Metrics: Monitor changes in customer satisfaction metrics.*   + *Operational Efficiency: Measure improvements in operational efficiency, especially in addressing pain points identified by the model.*   *Long-Term Impact:*   * + *Customer Retention: Anticipate improved customer retention due to enhanced service quality.*   + *Competitive Advantage: Gain a competitive edge by delivering personalized and satisfying experiences.*   *Implementation Cost:*  *Initial Implementation:*   * + *Data Science and IT Support: Estimate costs associated with hiring or contracting data scientists and IT professionals for initial implementation.*   *Ongoing Costs:*  *Staff Training Programs: Allocate resources for the development and delivery of customized staff training programs.*   * + *Monitoring and Maintenance: Budget for ongoing monitoring, refinement, and updates to the model.*   *Potential Savings:*   * + *Operational Efficiency: Identify potential cost savings through improved operational efficiency and reduced customer dissatisfaction issues.*   *ROI Assessment:*   * + *Regular Assessment: Periodically assess the ROI by comparing the implemented strategies' costs against improvements in customer satisfaction, revenue, and operational efficiency.*   *Key Considerations:*   * + *Flexibility: Be prepared to adapt the implementation plan based on real-time feedback and evolving business needs.*   + *Communication: Communicate changes, improvements, and outcomes to staff and customers to build trust and transparency.*   + *Continuous Improvement: Establish a culture of continuous improvement, with regular reviews and updates to the model and associated strategies.*   *By following this phased implementation plan, allocating resources strategically, and focusing on continuous improvement, the business can maximize the ROI of implementing the logistic regression model for predicting and enhancing customer satisfaction.* |

|  |
| --- |
| **References** |
| [*APA*](https://owl.english.purdue.edu/owl/resource/560/10/) *or* [*MLA*](https://owl.english.purdue.edu/owl/resource/747/08/) *format (10+)*  “About Us History | J.D. Power.” *Global Leader in Consumer Insights, Data & Analytics*, www.jdpower.com/business/about-us-history. Accessed 10 Dec. 2023.  Acharya, Akruti. “Training, Validation, Test Split for Machine Learning Datasets.” *Encord*, 13 June 2023, encord.com/blog/train-val-test-split/#:~:text=The%20rough%20standard%20for%20train,10%2D20%25%20test%20data.  Dhinakaran, Vigneshwer. “Exploratory Data Analysis (EDA) and Data Visualization with Python - Kite.” *Exploratory Data Analysis (EDA) and Data Visualization with Python*, 13 Nov. 2018, www.kite.com/blog/python/data-analysis-visualization-python/#:~:text=The%20correlation%20matrix,-A%20correlation%20matrix&text=This%20analysis%20allows%20you%20to,significant%20for%20building%20the%20model.  Jain, Akshay. “Advantages and Disadvantages of Logistic Regression in Machine Learning.” *Medium*, Medium, 22 July 2020, medium.com/@akshayjain\_757396/advantages-and-disadvantages-of-logistic-regression-in-machine-learning-a6a247e42b20.  K, Dhiraj. “Top 5 Advantages and Disadvantages of Decision Tree Algorithm.” *Medium*, Medium, 26 May 2019, dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a.  Patil, Prasad. “What Is Exploratory Data Analysis?” *Medium*, Towards Data Science, 30 May 2022, towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15.  Sequitin, Kirstie. “What Is an Outlier? Data Analytics Explained.” *CareerFoundry*, 28 Sept. 2023, careerfoundry.com/en/blog/data-analytics/what-is-an-outlier/#:~:text=In%20data%20analytics%2C%20outliers%20are,experimental%20errors%2C%20or%20a%20novelty.  Troy, Michael. “2023 North America Airline Satisfaction Study | J.D. Power.” *Airline Demand-Supply Imbalance Is Good for Revenue, Tough on Customer Experience, Says J.D. Power*, 10 May 2023, www.jdpower.com/business/press-releases/2023-north-america-airline-satisfaction-study.  “What Are Naive Bayes Classifiers?” *IBM*, www.ibm.com/topics/naive-bayes#:~:text=the%20next%20step-,Na%C3%AFve%20Bayes%20classifiers,a%20given%20class%20or%20category. Accessed 10 Dec. 2023.  “What Is Principal Component Analysis (PCA) and How It Is Used?” *Sartorius*, 18 Aug. 2020, www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186#:~:text=PCA%20forms%20the%20basis%20of,%2C%20jumps%2C%20clusters%20and%20outliers.  Wrobel, Melisa. “Data Normalization, Explained: What Is It, Why It’s Important, and How to Do It.” *InvGate*, InvGate Inc., 28 June 2023, blog.invgate.com/data-normalization. |

|  |
| --- |
| **Appendices** |
| *Exhibits (financials, market data, etc.)* |

|  |
| --- |
| **Sample Data Sources** |
| [UCI Data Sets](https://archive.ics.uci.edu/ml/index.php)  [MakeoverMonday Data Sets](https://www.makeovermonday.co.uk/data/)  [Kaggle Data Sets](https://www.kaggle.com/datasets)  https://www.kaggle.com/datasets/yakhyojon/customer-satisfaction-in-airline/data  [Amazon Web Services Public Data Sets](http://aws.amazon.com/datasets/)  [SEC Edgar Company Search](https://www.sec.gov/edgar/searchedgar/companysearch.html)  [Data.gov](http://www.data.gov/)  [Data Set Directory](http://blog.bigml.com/list-of-public-data-sources-fit-for-machine-learning/#open_companies)  [World Data Bank](http://databank.worldbank.org/data/home.aspx)  [Stetson Library](http://stetson.edu/library) |