SpaceX Intergalactic Teleportation Algorithm

Team Number 4
Jun Clemente, Tanya Ortega, Graham Ward
University of San Diego

Master of Science, Applied Data Science
ADS 501
June 24, 2024

Business Understanding

Background

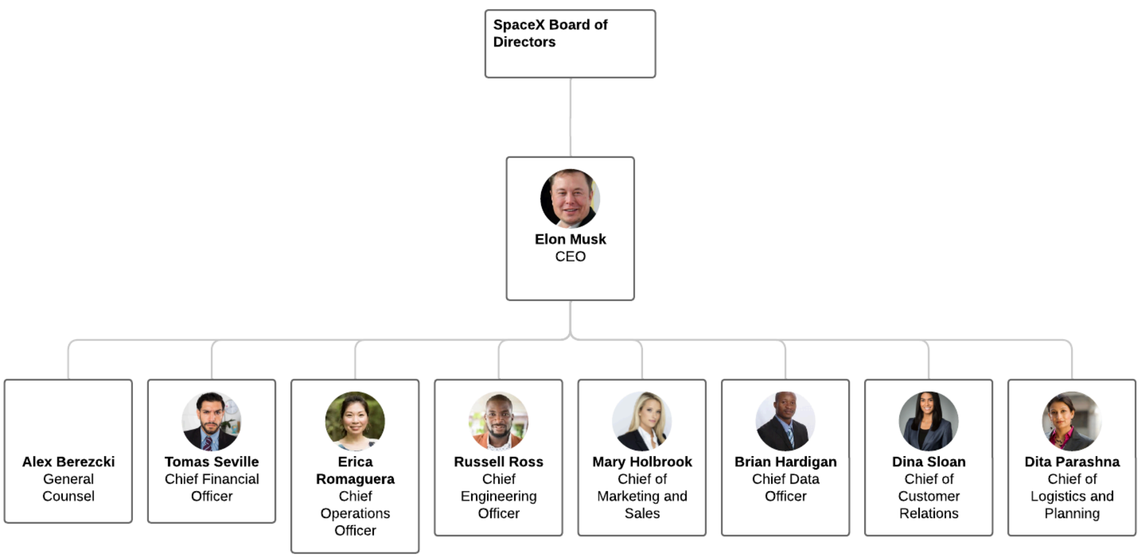
SpaceX Cruise Lines is a subsidiary of its parent company, SpaceX. SpaceX Cruise Lines was incorporated underneath SpaceX in the year 2850 and for the last 62 years has focused on developing interstellar space travel between habitable exoplanets, carrying passengers, supplies, and mail throughout the galaxy. Interplanetary travel throughout the universe has large untapped amounts of market value for SpaceX Cruise Lines as a leader in intergalactic travel. SpaceX has been leading the way in space exploration since early in the 21st century. A large portion of their success as a company comes from their continued leadership with Elon Musk as the CEO since the founding of SpaceX in 2002.

SpaceX's first interstellar cruise ship named *Spaceship Titanic* was on its maiden voyage traveling to the exoplanet 55 Cancri E. En route the cruise ship rounded the Alpha Centauri star system where it encountered and passed through a temporary space-time anomaly. This unexpected collision caused random passengers onboard to be transported to different dimensions. SpaceX successfully performed rescue missions for all of the passengers onboard that were transported by the temporary space-time anomaly. After the rescue mission, SpaceX conducted an Action Review (AAR), where they identified the need for a model that would provide real-time updates for passengers who were at higher risk of being transported and attempted to build a machine-learning algorithm to predict which passengers are susceptible to being transported via a temporary space-time anomaly based on data collected in real-time via data collected on passengers during their voyage. All data is maintained in the ship's mainframe, and remotely accessible on Earth via SpaceX's proprietary Warp Speed Data Relay Network(WS-DRN), with relay nodes positioned strategically throughout multiple galaxies.

In the event of another SpaceX Cruise Lines ship coming into contact with a temporal space-time anomaly, SpaceX Cruise Lines is attempting to build an algorithm that will correctly predict whether or not a passenger is susceptible to being transported to another dimension to support rapid deployment of rescue pods to the individual's dimensional coordinates.

Figure 1

Organizational flow chart for SpaceX Cruise Lines



Business Objectives and Success Criteria

Business Objectives: The primary goal of SpaceX Cruise Lines is to make intergalactic travel accessible to all, by providing affordable, comfortable, and safe luxury voyages across the reaches of the solar system. The goal of this project is to develop dedicated customizable rescue pods for passengers we predict will be teleported in the event of another anomaly. In order to accomplish this, SpaceX Cruise Lines is attempting to develop a predictive model to determine which passengers will be transported. Using this model, data will be fed to the rescue bay, where rescue pods will be prepared to be deployed as soon as any passenger's GPS tracking system leaves the dimensional coordinates of the ship.

Success Criteria: SpaceX Cruise Lines will define the success of the new algorithm from a business perspective by having 95% of all transported passengers returned to the ship within 1 minute of being transported and the remaining 5% within 5 minutes of being transported. The remaining 5% of the rescue pods will all be deployed with tactical teams positioned on standby due to the random nature of the teleportation locations in order to ensure 100% rescue within 5 minutes. The technical success of this model will be evaluated using an AUC-ROC curve, which has an AUC value above 0.95, indicating the model correctly predicts who will and who will not be subject to interdimensional travel 95% of the time.

Inventory of Resources

Due to the potential effects of another space-time anomaly on the current quantum computer systems in use, the hardware for this project is a custom-built desktop computer using parts sourced from the 21st century. The PC consists of an AMD Ryzen 7 3700X 3.6 GHz 8-Core processor, 64 GB of RAM, and runs Microsoft Windows 11 as the operating system. The software was selected from the historical archives to ensure compatibility with the PC for analyzing and processing the provided data. This includes Microsoft Visual Studio Code, Microsoft Office 365 Excel and Word, Google Colab, Google Drive, GitHub, GitHub Desktop, and

WinZip. The information on the passengers was collected and provided by Kaggle.com (*Spaceship Titanic | Kaggle*, n.d.). This will serve as the base for the analytics base table (ABT) for this mission. Based on the requirements for this mission and outlined in this proposal, the software and hardware listed will fully support the preparation, building, and running of the proposed predictive model.

Requirements

In order for this project to be successful, there are some key requirements surrounding the schedule of completion, quality of the results, and legal conditions which will be addressed here. Based on the existing infrastructure, our current fiscal budget, and the need to develop these systems before any more cruise liners make their way into the market this project will have an estimated completion time of nine full weeks. These nine weeks will be broken into three distinct phases for the predictive analytics team. During these nine weeks, while the predictive analytics team is performing their work, the engineering division combined with the safety department will be responsible for developing the rescue pod program and ensuring the safety of each and every passenger and crew aboard every ship.

In addition to the proposed nine-week timeline, the quality of the results will speak to the success of the predictive algorithm as a whole. The model is expected to have all relevant statistical measures such as a high Adjusted R-squared, no instances of multicollinearity, and most importantly the metric used to evaluate the success of the model will be an Area Under Curve - Receiver Operating Characteristics (AUC-ROC) that has a value of AUC = 0.95. This indicates the model would correctly be able to distinguish between the classes of transported and not transported with 95% accuracy (Narkhede, 2018). Using the AUC-ROC curve allows us to gain insight into the True Positive Rate, those that were predicted to be transported and were; as well as the True Negative Rate, those that were predicted not to be transported and were not. By setting the AUC equal to 0.95 we are able to examine the underlying features of why some people are transferred and others are not, which in turn gives us deeper insight into the science of these space-time anomalies.

Since the data being used in this project is property of SpaceX Cruise Lines the use of this dataset in its current form complies with all current planetary laws and regulations as well as intergalactic. However, if the model is to be adapted for future use, supplemented with additional features, or retrained on new data then SpaceX Cruise Lines will make sure to act in accordance with data and privacy laws established for the user's home planet. SpaceX is committed to data privacy and personal data for this project and all future projects will only be used to develop better products and features for SpaceX passengers and will never be sold to third-party vendors.

Assumptions

All data used in this project has been collected from SpaceX Cruise Lines’ onboard computer tracking software and stored in the passenger database property of SpaceX Cruise Lines. All passenger data that SpaceX Cruise Lines collects is in accordance with each user's home planet data privacy requirements. All technology is assumed to operate as it did before entering the space-time anomaly and operates under the assumption technology will continue to function as intended after the transportation has taken place. In addition, this project assumes that no changes in regulatory policy are expected to come down from any planet or galactic governing body within the next Earth year that would inhibit the use of this model in anything but its planned and established operation.

Constraints

SpaceX Cruise Lines is currently operating under constraints with respect to the predictive analytics model as well as the physical rescue pod program that will be supported by the predictive analytics model. In the predictive analytics model, the biggest constraint is the dataset. Currently the dataset represents just one event of a ship passing through a space-time anomaly and its effects on the passengers. This understanding and data could change as we learn more and more about the space-time anomalies. Another important limitation of the dataset is that there is only data about the customers during their time onboard the ship. SpaceX Cruise Lines is hoping to remove this constraint by enhancing the dataset with new features, and slowly incorporating more and more details about passengers, crew, and incorporating new data into the model for further testing and analysis. This will allow SpaceX Cruise Lines to identify different features responsible for certain passengers and crew members being transported. With respect to data transmissions, this project is constrained in the sense that it relies on technology operating in good working condition, with multiple moving parts, spread across the universe in order to quickly locate, identify, and retrieve passengers and crew no matter where they end up getting transported to. This means if any technology throughout the SpaceX ecosystem is not maintained then it has the potential to inhibit any rescue operations and constrain us from employing our rescue technology in the most efficient and effective manner possible.

From the standpoint of implementing the rescue pod operation, constraints exist pertaining to the number of rescue crew members that will be able to fit aboard the craft. Furthermore contingencies must be put in place for what happens if the rescue crews themselves are part of the transported personnel. Furthermore, rescue pod operations are constrained by the need to be manually activated for the 5% of the personnel who were not successfully returned to the ship within the first minute. This forced increase in the number of rescue personnel requires us to either make the ships bigger, alter the interior design to accommodate more crew, or limit the number of rescue personnel brought on each voyage, which runs the risk of not being able to meet our goals for this project.

RESOLVEDD Strategy (Chapman et al., 2000)

During a voyage, passengers on the cruiseliner encountered a space-time anomaly that transported some passengers to another dimension. As a result, these passengers needed to be rescued from their new location. In the event of unexpected dimensional passenger transportation, the company must preemptively know which passengers will be transported in the event of a space-time anomaly. By analyzing historical data through predictive data analysis, it is possible to develop an algorithm to identify the factors that result in these anomalies.

Without any changes, the company will continue to bleed high costs for these rescue missions without improving passenger experience. On the other hand, if predictive data analysis is implemented, amenities and other features that contribute to dimensional transportation can be monitored. This proactive approach would result in faster rescue missions and achieving the goal of 95% retrieval within the first minute of transport.

Prioritizing passenger safety and reducing liabilities will improve the company's reputation and financial performance. Through implementation of these measures, stock prices could increase and the company could be known as the safest travel option in the industry. This approach aligns with the company's core values and long-term goals.

There are two solutions moving forward. One solution is to maintain the status quo and continue as is and perform rescue missions through the current protocol. The other solution is to use predictive data analysis to develop an algorithm that identifies passengers that will be transported in the event of a space-time anomaly, therefore improving rescue times to one minute within transport. Additionally, this algorithm can be used to determine which amenities and features result in dimensional transportation allowing for changes to be made to mitigate and eliminate transportation and therefore reducing the budget for rescue missions. This approach will enhance passenger safety, reduce retrieval costs, and improve the overall satisfaction of our customers.

There are a few potential challenges to this strategy such as ensuring the accuracy of the data used for predictive analysis. Buy-in from the board of directors and C-level executives may also be difficult and would need to be convinced of the effectiveness and cost benefits of our solution. Additionally, changes to amenities and features could result in negative customer satisfaction in the short term. Nevertheless, the company's values of customer importance and satisfaction will be upheld. Additionally, with this proactive approach, the company's goal of being the safest travel option in the industry will also improve their value and stock prices.

Risks And Contingencies

SpaceX Cruise Lines maintains a deep commitment to safety and in doing so, identified risks in both the rescue operation project and the data used in the predictive analytics model and developed contingencies in support of their overarching goal to keep passengers safe. On the rescue operation side of the project, risks include incorrect rescue coordinates or none at all, potential loss of life during transportation or within the five-minute window, inadequate number of rescue pods, and insufficient rescue crews for worst-case scenarios. With regards to the predictive analytics model, risks include a lack of thorough understanding of the science behind space-time anomalies, the need to account for concept drift as further data is collected and our understanding grows, and finally whole data manipulation or eradication caused by another event.

By developing contingencies to these risks, SpaceX aims to ensure that intergalactic travel is safe and affordable for all throughout the universe. SpaceX developed contingency plans addressing the risks of both the rescue pod operations as well as the data risks associated with the predictive analytics model. Regarding the rescue pod operation, each passenger on the vessel will be given two bracelets that will act as dual protective barriers from warping, unfriendly atmospheric gasses and living conditions, as well as allowing breathing in the vacuum of space. In the unlikely event that there is a casualty during an event, all rescue crews will carry recovery kits to return remains to next of kin. The fears about not having enough rescue pods currently on the fleet necessitated a complete redesign of the ship's hull to accommodate new regulations on a mandatory one-to-one rescue pod-to-ship passenger ratio, to include crew members. Finally addressing the issue of lack of adequate rescue personnel in the event of a worse case scenario, each cruise line will travel in tandem with a rescue support vessel to aid in mass rescue events.

From the perspective of the data analytics model, most of the risks come from our current lack of understanding of the inner workings of space-time anomalies. Risks needing contingency planning for the model involve our data not being an accurate representation of the science behind these anomalies, the need to address concept drift as the model learns from future events and intakes new data on persons transported, and finally the overall quality and reliability of data relayed at and around the time of the event. In order to combat these risks, SpaceX Cruise Lines developed contingencies for each of these. The first two risks are similar and addressed in the same way. Both the lack of understanding about the science behind why these anomalies don't affect everyone and needing to account for concept drift can be solved by developing a model designed to continually learn from these events. The initial model will be robust, containing data from the ships logs as a starting point for the model and then adding the more personal data points such as race, body morphology, overall level of health, time spent in space, and more. The issue of concept drift can be addressed through the constant transmission of data through the WS-DRN. This constant data flow will keep the models updated on real time data including all the financial and ship data, but also biometric data that will be continually scanned and transmitted. This will allow the model to update in near real time, allowing for the updating of persons that are at higher risk for being transported in the case of an event. Addressing the final risk of overall data quality being affected by an event we again utilize the proprietary WS-DRN to ensure a constant stream of data on all of the passengers and crew members, which will allow for us to identify when data has been corrupted by comparing it to the data that was relayed just beforehand. At each relay node in the WS-DRN calculations are performed on the accuracy of the data transmitted from the previous node in the network, which will allow us to monitor the transmission quality of the data as well. By taking advantage of SpaceX's data relay network we will be able to ensure the highest quality of the data coming in and be able to discard corrupted data, so that the model does not learn from dirty or corrupted data.

Terminology

The following terms are essential for understanding interstellar space travel and the context of SpaceX Cruise Lines. Each term introduces a crucial role in shaping the analysis and interpretation of the data.

Action Review (AAR): A process to analyze the performance of a mission after its completion and areas to improve.

AUC-ROC: A performance measurement for classification models, it indicates the true positive rate vs the false positive rate. Acronym for Area Under the Curve - Receiver Operating Characteristics.

Concept Drift: A change in the statistical properties of the target variable over time.

Covariance and Correlation Analysis: A Statistical method used to measure the relationship between two variables.

Cruise lines: Companies providing passenger ships for travel with intentions of leisure travel or voyages lasting more than one day and various destinations. In this case, SpaceX applies the interstellar voyage.

Cryosleep: When a human's body temperature is lowered and slows the biological process in order to safely travel for a long duration.

Dimensional Coordinates: Four-dimensional coordinate system that allocates for all three

Dimensional Transportation: The process of transportation from one dimension to another.

Exoplanets: Planets orbiting stars outside the solar system.

Interstellar: In this context, it refers to travel in space between star systems.

Machine Learning Algorithms: enable computers to learn from data and make predictions or decisions.

Nearest Neighbor Algorithm: a type of machine learning algorithm that builds a model on the measures of similarity.

Predictive Model: A model used to predict future events by using historical data. The model aims to predict if a passenger is being transported by a space-time anomaly.

Space-time Anomaly: Irregularity in the space-time, potentially causing unexpected events or phenomena.

Transported: If the passenger was transported by space-time anomaly during the voyage.

VIP Status: A designation for passengers receiving special privileges or services.

Warp Speed Data Relay Network (WS-DRN): SpaceX's proprietary network for transmitting data across vast interstellar distances.

Data Mining Goals and Success Criteria

The intended output of this project is to develop a predictive analysis algorithm that accurately predicts the probability a passenger will be transported to an alternate dimension during a space-time anomaly. Historical data from the *Spaceship Titanic's* internal logbook will be used for this analysis. Using data exploration and similarity-based machine learning techniques, the data science team will develop an algorithm to identify the features that lead to passenger transportation during unforeseen space-time anomalies.

Data mining and exploration of the data will provide insight into which variables are correlated with passengers being transported. Data quality analysis will be performed to determine the viability of each feature in the dataset. Any features missing more than 60% of their values will be removed from the dataset (Kelleher et al., 2020). For the remaining features, imputation will be used to fill in missing values. Covariance and correlation analysis will be performed on both continuous and categorical data to help with feature selection.

The success criteria for this project is to develop a highly accurate predictive model nearing 100% accuracy. Although a model that is 100% accurate may be unrealistic due to the complexities in predictive analysis, an AUC-ROC of >= 0.95 will be considered a successful model. After the initial exploratory data analysis, predictive analysis modeling will be performed using *nearest neighbor algorithms*. Feature selection and retraining of the model will be repeated until a model that satisfies the success criteria is developed.

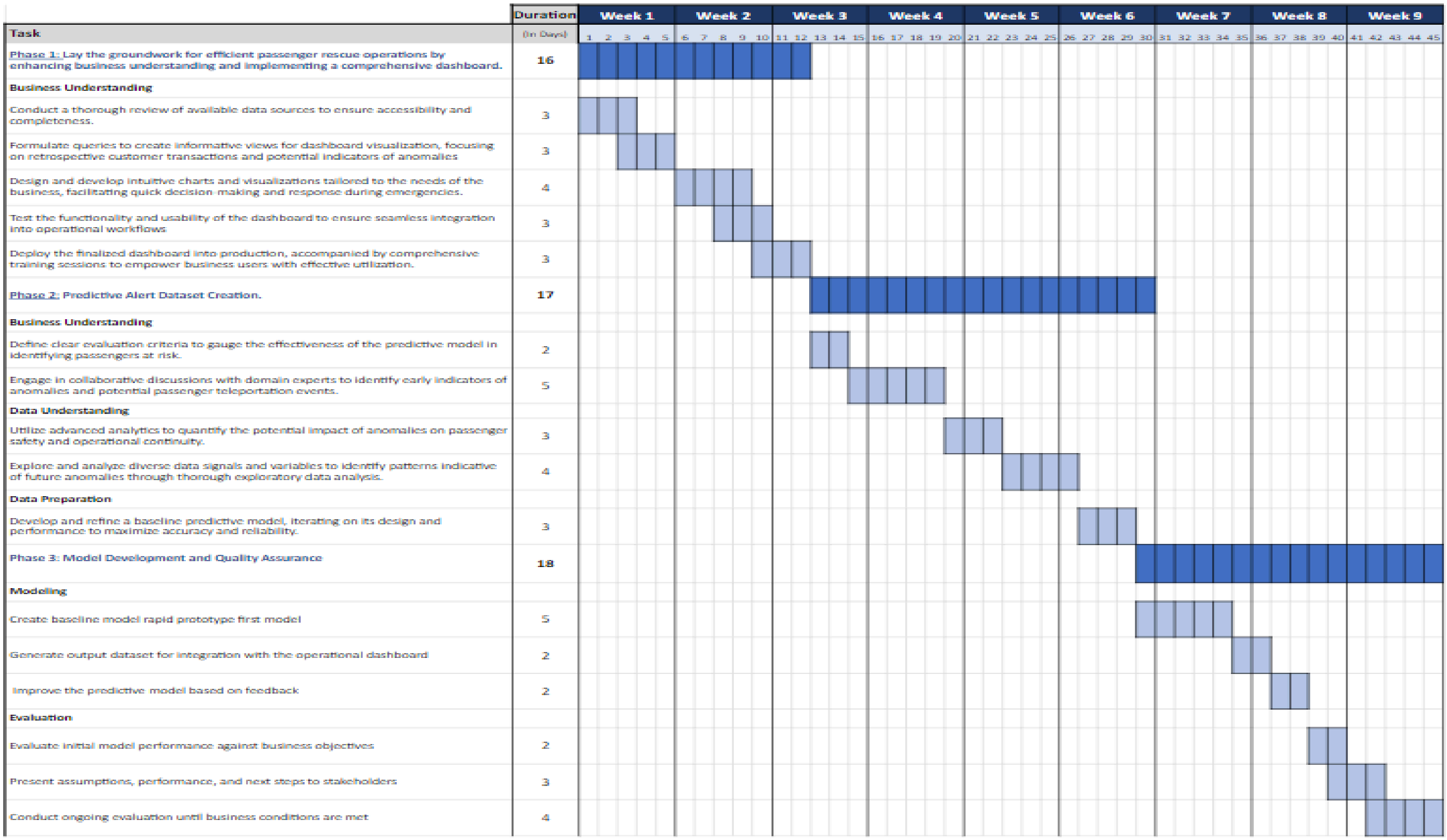
Project Plan/ Order of Tasks

The project plan aims to enhance passenger space operations by implementing a comprehensive dashboard through a three-phase approach over ten weeks. The first phase focuses on understanding the business, formulating queries, designing visualizations, and integrating the dashboard while training employees for effective use. The second phase involves creating a predictive alert dataset by defining evaluation criteria, engaging with domain experts, analyzing data for anomalies, and developing a baseline predictive model.

In the third phase, the project shifts to model development and quality assurance, where a rapid prototype of the predictive model is created, integrated with the operational dashboard, and improved based on feedback. The final evaluation phase involves assessing the model's performance against business objectives, presenting findings to stakeholders, and conducting ongoing evaluations to ensure the model meets the required standards. This structured plan ensures systematic progress and improves passenger operations.

Figure 2

Gantt Chart Project Timeline



Data Understanding

Initial Data Collection Report

The initial data collected for this predictive model is taken from the *Spaceship Titanic's* internal electronic logbook. This data contains passenger information for the 12,972 passengers that were on board the intergalactic voyage. The data collected from the ship's mainframe computer contains information on each passenger such as the unique passenger identification number, what planet they are coming from, their destination planet, whether or not they chose to be suspended in cryo sleep, their VIP status, their specific room number, their age, their name, the amount each passenger has spent at different departments on the cruise, and finally the variable we are attempting to predict, whether or not the passenger would be transported in the event of another collision with a temporary space-time anomaly**.**

Data Description Report

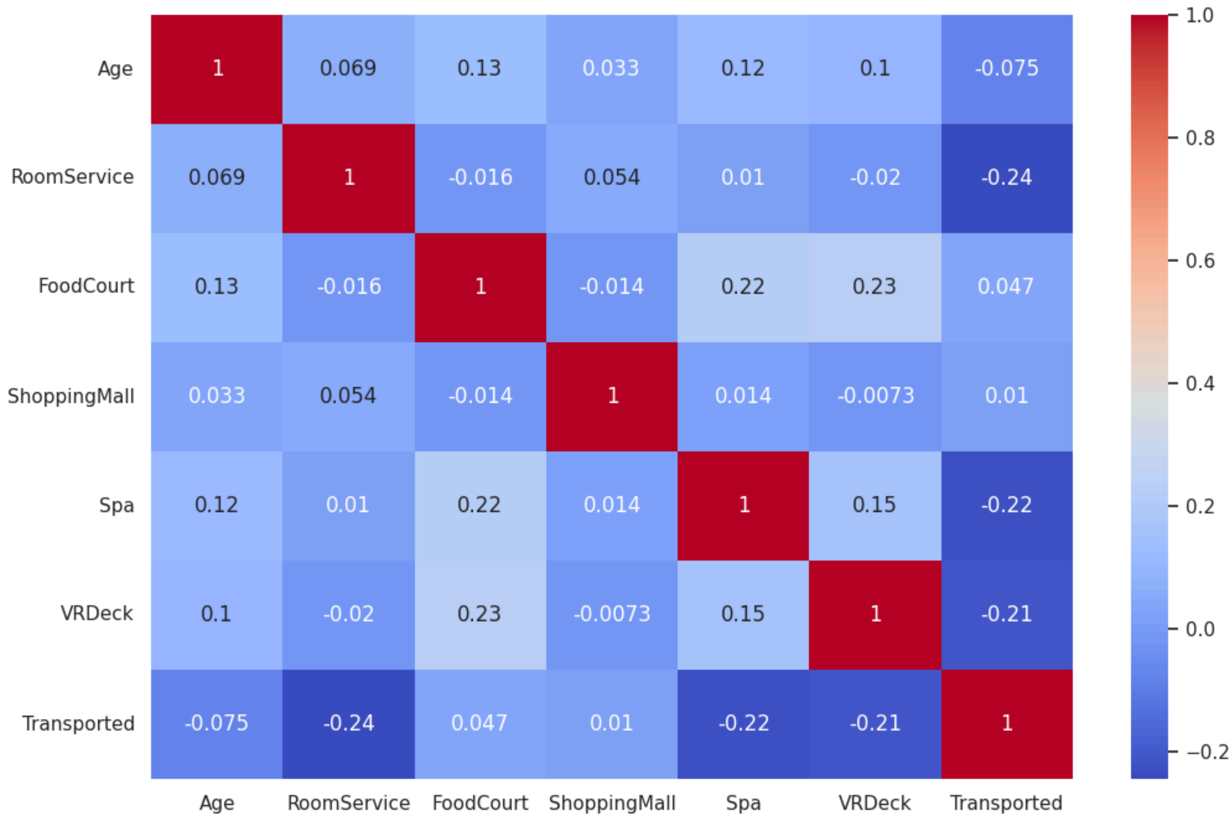
File 1: Spaceship-Titanic-Train.csv

A dataset has been collected from the Spaceship Titanic, operated by SpaceX Cruise Lines for space travel. This dataset contains information about the passengers who are on the voyage. The dataset holds information for 8,694 passengers, of which all numbers are slightly different and require more clarification. Each passenger record includes details about their home plant, destination planet, and if they chose to be cryogenically frozen during their voyage. The dataset also includes the passenger's assigned cabin, their age at the time of departure, and VIP status if applicable. While passengers are on their voyage, they also track how much passengers spend in onboard departments on the spaceship such as; room service, food court, shopping mall, spa, and VR deck. Finally, the dataset includes a crucial column for the machine-learning model: a binary variable if the passenger was transported by space-time anomaly during the voyage labeled as "Transported" (Yes/No), which will be used as an essential variable for the predictive model.

Data Exploration Report

Figure 3

Heatmap Correlation Matrix

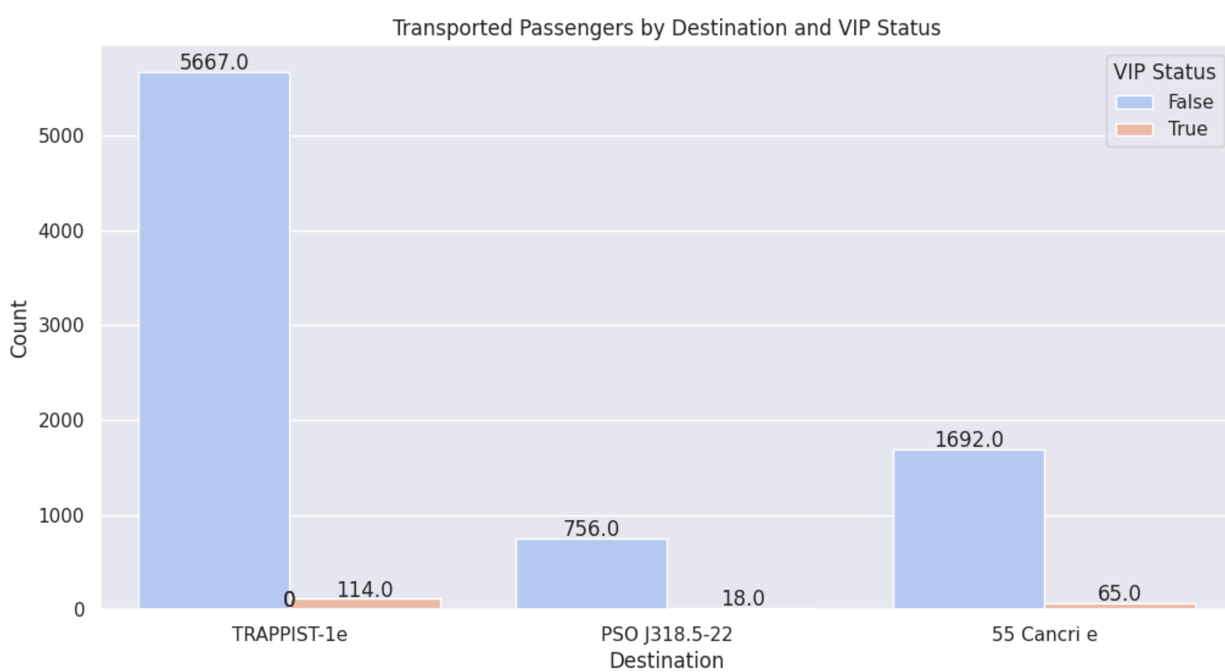


This heatmap shows the correlation between various passenger attributes (Age, RoomService, FoodCourt, ShoppingMall, Spa, VRDeck) and whether they were transported (Transported). The color scale indicates the strength and direction of the correlations, with red representing positive correlations and blue representing negative correlations. Therefore, RoomService (-0.24), Spa (-0.22), and VRDeck (-0.21) have moderate negative correlations with being transported, suggesting that higher usage of these services is associated with a lower likelihood of transportation.

On the contrary, FoodCourt has a weak positive correlation (0.047) with being transported, indicating that passengers using the FoodCourt are slightly more likely to be transported. Other variables, such as Age and ShoppingMall, show very weak correlations with transportation status, indicating that these factors have minimal impact on whether a passenger is transported. This analysis can help identify key factors affecting passenger transportation outcomes.

Figure 4

Bar Chart of Transported Passengers by Destination Planet and VIP Status

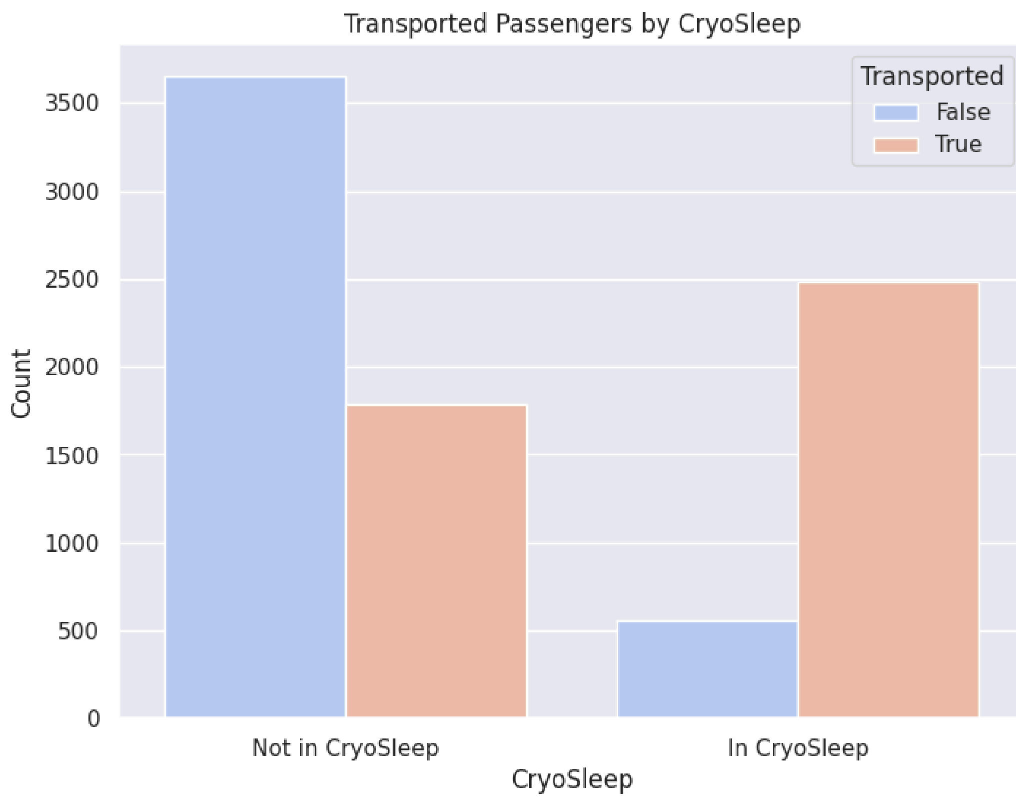


The chart depicts the number of passengers transported by SpaceX Cruise Lines to three destinations: TRAPPIST-1e, PSO J318.5-22, and 55 Cancri e, categorized by VIP status. The data shows a significant dominance of non-VIP passengers (blue) over VIP passengers (orange) across all destinations. TRAPPIST-1e has the highest number of passengers, with 5667 non-VIPs and 114 VIPs. 55 Cancri e follows with 1692 non-VIPs and 65 VIPs, while PSO J318.5-22 has 756 non-VIPs and 18 VIPs.

This distribution indicates that TRAPPIST-1e is the most popular destination, especially among non-VIP passengers. The chart highlights a potential opportunity for SpaceX Cruise Lines to enhance services or marketing strategies targeted at VIP passengers, given their relatively lower numbers across all destinations. It is more than likely that passengers who are VIPs have a lower chance of being transported. Also, the data appears to have the most passengers in TRAPPIST-1e which may have attributed to several factors. The popularity may appeal to passengers due to unique experiences making it a desirable destination. Paying close attention to why passengers are choosing TRAPPIST-1e can provide valuable insights into increasing the number of passengers to other destinations. If passengers prefer TRAPPIST-1e due to desirable factors, leveraging these insights can enhance marketing strategies and improve business operations. This approach can help achieve a better balance of passengers across multiple destinations and ensure safer travel experiences.

Figure 5

Transported Passengers in CryoSleep



The chart shows the number of passengers transported to different dimensions based on their CryoSleep status, divided into those not in CryoSleep and those in CryoSleep, and further categorized by whether they were transported (orange) or not (blue). Among passengers not in CryoSleep, approximately 3500 were not transported while around 2000 were transported. For those in CryoSleep, about 500 were not transported, whereas approximately 2500 were transported. This suggests that passengers in CryoSleep have a higher likelihood of being transported to different dimensions compared to those not in CryoSleep.

Data Quality Report

A review of the provided dataset revealed that there were six features of continuous variables (Table 1) and seven features of categorical variables (Table 2). The target attribute is the “Transported” variable. Cardinality was reviewed for each feature for any potential errors. Descriptive statistics were performed on all variables. The total missing values for each feature was also calculated showing that all columns were missing <3.0% of values.

Through data quality analysis of the dataset, all features are viable for use in predictive analysis. It was calculated that none of the features are missing 60% or more of their data. Therefore, all features will be retained. To ensure the best performance for machine learning, missing values for continuous features will be imputed with the median. For categorical features, the missing values will be imputed with the feature mode.

Table 1

Data Quality Report Table for Continuous Features

Feature	**Count**	**% Miss.**	**Card.**	**Min.**	**1st Qrt.**	**Mean**	**Median**	**3rd Qrt.**	**Max.**	**Std. Dev.**
Age	8514	2.10	81	0.00	19.00	28.83	27.00	38.00	79.00	14.49
Room Service	8512	2.13	1274	0.00	0.00	224.69	0.00	47.00	14327.00	666.72
FoodCourt	8510	2.15	1508	0.00	0.00	458.08	0.00	76.00	29813.00	1611.49
Shopping Mall	8485	2.45	1116	0.00	0.00	173.73	0.00	27.00	23492.00	604.70
Spa	8510	2.15	1328	0.00	0.00	311.14	0.00	59.00	22408.00	1136.71
VRDeck	8505	2.21	1307	0.00	0.00	304.85	0.00	46.00	24133.00	1145.72

Table 2

Data Quality Report Table for Categorical Features

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
PassengerID	8693	0.00	8693	0001_01	1	0.01	6136_01	1	0.01
HomePlanet	8492	2.37	4	Earth	4602	54.19	Europa	2131	25.09
CryoSleep	8476	2.56	3	False	5439	64.17	True	3037	35.83
Cabin	8494	2.34	6561	G/734/S	8	0.09	G/109/P	7	0.08
Destination	8511	2.14	4	TRAPPIST-1e	5915	69.50	55 Cancri e	1800	21.15
VIP	8490	2.39	3	False	8291	97.66	True	199	2.34
Name	8493	2.35	8474	Gollux Reedall	2	0.02	Elaney Webstephrey	2	0.02
Transported (Target Var)	8693	0.00	2	True	4378	50.36	False	4315	49.64

Table 3

Data Quality Plan

Feature	Data Quality Issue	Potential Handling Strategies
Continuous features	Missing values	Impute missing values with the feature median
Categorical features	Missing values	Impute missing values with the feature mode

References

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*. https://mineracaodedados.wordpress.com/wp-content/uploads/2012/12/crisp-dm-1-0.pdf

Evidently AI Team. (n.d.). *What is concept drift in ML, and how to detect and address it*. Www.evidentlyai.com. https://www.evidentlyai.com/ml-in-production/concept-drift

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2020). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies* (Second edition). MIT Press.

Narkhede, S. (2018, June 27). *Understanding AUC - ROC Curve*. Medium; Towards Data Science. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Spaceship Titanic | Kaggle. (n.d.). Retrieved May 27, 2024, from https://www.kaggle.com/competitions/spaceship-titanic/data