Semester Project

Dohyun Kim, Kaehyun Um[1]*

Abstract
Spaceship Titanic: Predict which passengers are transported to an alternate dimension
House Prices: Predict Sales prices

Keywords
EDA — Predict — Regression

[1] *Computer Science, School of Informatics , Computing and Engineering, Indiana University, Bloomington, IN, USA*

Contents

1. Problem and Data Description

Spaceship Titanic: This problem assumes that a spacecraft, named Titanic, carrying passengers has a problem while sailing in the very distant future. With this data set, we need to figure out which passengers were transported by the anomaly.

Read data : train.csv

- PassengerId - A unique Id for each passenger. Each Id takes the form where indicates a group the passenger is travelling with and pp is their number within the group. People in a group are often family members, but not always.
- HomePlanet - The planet the passenger departed from, typically their planet of permanent residence.
- CryoSleep - Indicates whether the passenger elected to be put into suspended animation for the duration of the voyage. Passengers in cryosleep are confined to their cabins.
- Cabin - The cabin number where the passenger is staying. Takes the form deck/num/side, where side can be either P for Port or S for Starboard.
- Destination - The planet the passenger will be debarking to.
- Age - The age of the passenger.
- VIP - Whether the passenger has paid for special VIP service during the voyage.
- RoomService, FoodCourt, ShoppingMall, Spa, VRDeck - Amount the passenger has billed at each of the Spaceship Titanic's many luxury amenities.
- Name - The first and last names of the passenger.
- Transported - Whether the passenger was transported to another dimension. This is the target

House Prices: The problem with this data is predicting the sales price of each house. You can predict the value of the SalePrice for each ID. This data contains many things that homebuyers are curious about. For example, the quality and area of the garage were built in what year.

Read data: train.csv

- SalePrice: the property's sale price in dollars. This is the target variable that you're trying to predict.
- LotArea: Lot size in square feet
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- ExterCond: Present condition of the material on the exterior
- BsmtCond: General condition of the basement
- HeatingQC: Heating quality and condition
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- GarageCond: Garage condition

2. Data Preprocessing & Exploratory Data Analysis

2.1 Handling Missing Values

Spaceship Titanic

We will just drop the missing values of HomePlanet, CryoSleep, Cabin, Destination, VIP, and Name to 0. Because if there is no value, it means that this data is not important for analysis. But for Age, RoomService, FoodCourt, ShoppingMall, Spa, and VRDeck, we would replace missing values with the average of each variables.

House Prices

We will just drop every NA data because most of the data is categorical data.
There are 81 columns in the data. And we found that there are so many NaN value which is unnecessary while we explore the data. So we decided to drop the value which has so many NaN value.

2.2 Exploratory Data Analysis

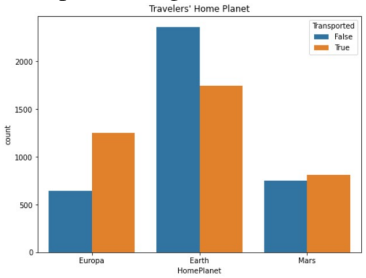
Spaceship Titanic

(1) Overview the data

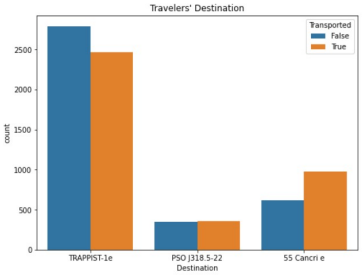
	mean	std	min	25%	50%	75%	max
Age	28.846256	14.333877	0.000000	20.000000	27.000000	37.000000	79.000000
RoomService	221.319026	631.582409	0.000000	0.000000	0.000000	80.000000	9920.000000
FoodCourt	464.970397	1626.530777	0.000000	0.000000	0.000000	122.000000	29813.000000
ShoppingMall	177.069542	561.472743	0.000000	0.000000	0.000000	47.000000	12253.000000
Spa	308.882566	1117.335508	0.000000	0.000000	0.000000	92.000000	22408.000000
VRDeck	305.237307	1116.902993	0.000000	0.000000	0.000000	74.500000	20336.000000

	count	unique	top	freq
PassengerId	7559	7559	0001_01	1
HomePlanet	7559	3	Earth	4101
CryoSleep	7559	2	False	4854
Cabin	7559	5957	G/734/S	8
Destination	7559	3	TRAPPIST-1e	5253
VIP	7559	2	False	7379
Name	7559	7540	Glena Hahnstonsen	2

(2) Explore Categorical Data

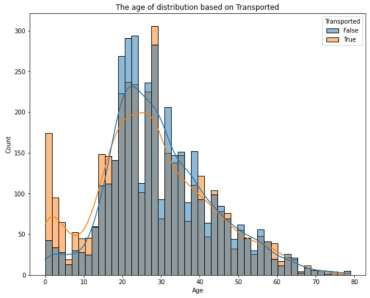


Description: Based on transported, we used a count plot to show where the passengers' HomePlanet is.



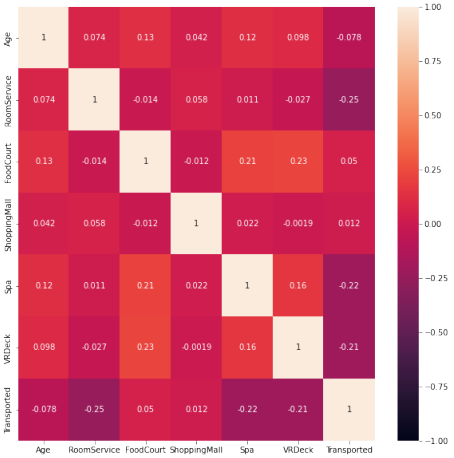
Description: Based on transported, we used a count plot to indicate where the guests are headed.

(3) Explore Continuous Data



Description: Graphed the age distribution based on transported.

(4) Check the correlation between columns

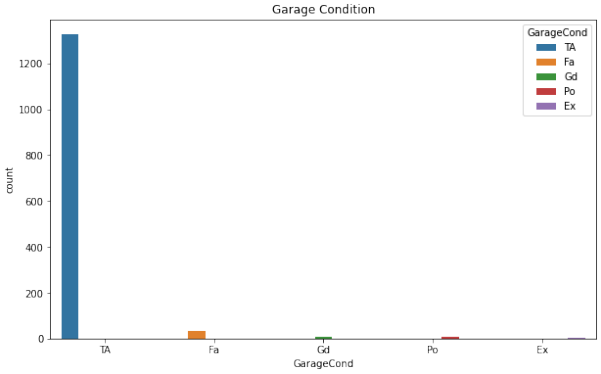


House Prices

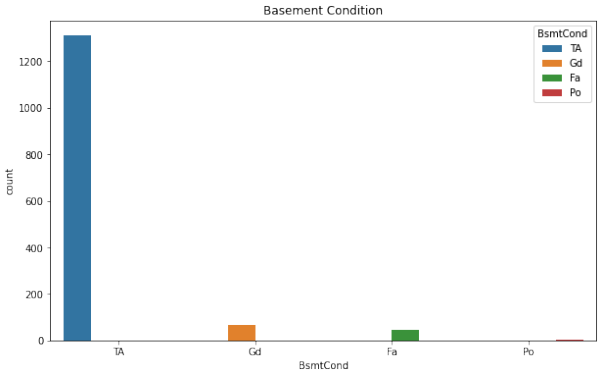
(1) Overview the data

	mean	std	min	25%	50%	75%	max
SalePrice	180921.195890	79442.502883	34900.000000	129975.000000	163000.000000	214000.000000	755000.000000
LotArea	10516.828082	9981.264932	1300.000000	7553.500000	9478.500000	11601.500000	215245.000000
OverallCond	5.575342	1.112799	1.000000	5.000000	5.000000	6.000000	9.000000
YearBuilt	1971.267808	30.202904	1872.000000	1954.000000	1973.000000	2000.000000	2010.000000
YearRemodAdd	1984.865753	20.645407	1950.000000	1967.000000	1994.000000	2004.000000	2010.000000
1stFlrSF	1162.626712	386.587738	334.000000	882.000000	1087.000000	1391.250000	4692.000000
2ndFlrSF	346.992466	436.528436	0.000000	0.000000	0.000000	728.000000	2065.000000

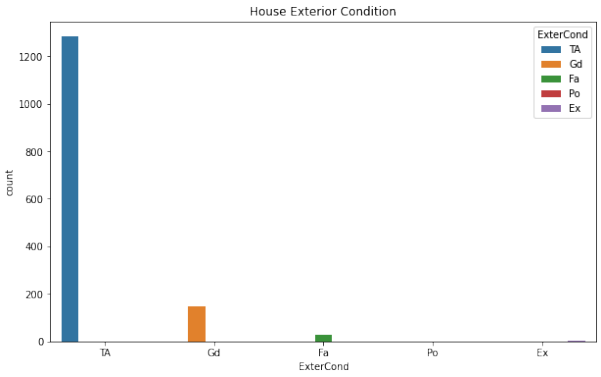
(2) Explore Categorical Data



Description: For each column named GarageCond in the house price data, the count plot is used to show a graph of Garage Condition.

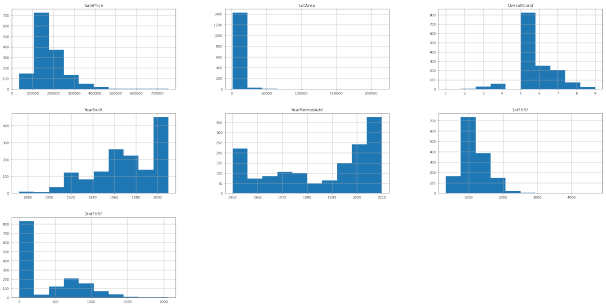


Description: For each column named BsmtCond in the house price data, a count plot is used to represent the graph.



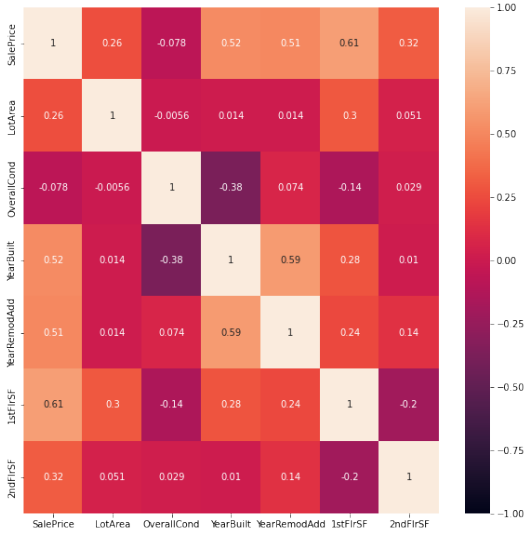
Description: For each column named ExterCond in house price data, ExterCond is represented using count plot.

(3) Explore Continuous Data



Description: The successive data of the house price data were plotted as a graph using a histogram.

(4) Check the correlation between columns



3. Algorithm and Methodology

Spaceship Titanic

1. Preprocess the data

In order to process the data for the model, the data was reloaded and the missing values of various columns were filled using the media method using Simple Imputer. And the Homeplanet column used the mode method to remove the missing values. Also, we converted categorical data using label encoder. Finally, using the train_test_split function, we separated the training set and the test data set and prepared the data to be modeled.

2. Introducing the Model Used

Summary : Overall, we trained 6 models and got predictions. In particular, we applied the GridSearchCV module to find the parameters that can show the highest accuracy for each model. In addition, the best_score_ of GridSearchCV was used to derive the optimal accuracy.

Model (1): K-Nearest Neigbors (KNN)
Accuracy: 68.00%

KNN is an algorithm of selecting and predicting K-nearest classes, and Euclidean distance is mainly used as a metric. Although it is a simple method, the results can be sensitive by various factors such as the value of k, preprocessing, and metrics.

Model (2): Gaussian Naive Bayes
Accuracy: 70.00%

Naive Bayes classifies samples by using Bayes' Theorem. Especially, Gaussian Naive Bayes is used when the data are continuous, and the probability is calculated under the assumption that the values are distributed according to a normal distribution.

Model (3): Logistic Regression
Accuracy: 78.00%

Logistic Regression is a supervised learning algorithm that uses regression to predict the probability of data between 0 and 1 and classify them as being in a more likely class based on that probability.

- Ensemble Methods

Model (4): Random Forest
Accuracy: 78.00%

Random Forest(RF) is a kind of bagging method, which works by outputting regression from a number of decision trees configured during the training process. The factors used to make each decision tree are randomly selected. It is useful to deal with missing values and also works well with both categorical and continuous data.

Model (5): XGBoost
Accuracy:77.00%

XGBoost(XGB) is one of boosting methods. It is similar to Random Forest in that it is consisted of an ensemble of decision trees. There is a difference in how the trees are derived. XGB uses extreme gradient. It performs tasks faster than existing GBMs (Gradient Boosting Methods).

Model (6): LightGBM
Accuracy: 80.00%

Light GBM(LGBM) is basically the same as XGB, but works with a different boosting technique. This algorithm uses the leaf-centered tree segmentation method. It usually produces results similar to XGB, but is much faster and uses less memory.

House Prices

1. Preprocess the data

In order to process the data for the model, the data was reloaded and the missing values of various columns were filled using the media method using Simple Imputer. And the other columns filled in the missing values for each situation. In addition, the label encoder was used to transform categorical data accordingly. We also standardized the data using the StandardScarler function. Finally, using the train_test_split function, we separated the training set and the test data set and prepared the data to be modeled.

2. Introducing the Model Used

Summary: Overall, we trained the models and got the predicted values. To evaluate the performance of the model, we used r2_score that shows the coefficient of determination.

Model (1) : Linear Regression
Fitness assessment : 0.849

Linear Regression is a method to model the relationship between variables as a linear function expression on the premise that independent variables and dependent variables are linearly related.

Model (2) : Lasso
Fitness assessment : 0.849

The Lasso model is an alternative to the Ridge model, where all elements of the weight are zeroed or close to zero. In short, a linear regression method using L1 normalization. When using the Lasso model, the regulation coefficient alpha was applied as 0.01.

Model (3) : XGBoost Regressor
Fitness assessment : 0.930

The XGBoost Regressor model uses the boosting method, if applicable to the ensemble technique. This model is designed to overcome the computational limitations of boost tree algorithms. This model is often used as a classifier, but it also has a good performance, so I used it to analyze this data.

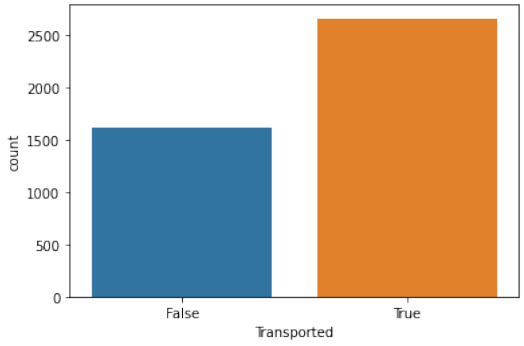
Model (4) : LightGBM Regressor
Fitness assessment : 0.910

The LightGBM Regressor model corresponds to the boosting of the ensemble technique. It is a model that compensates for the shortcomings of XGBoost, which is much faster and less memory usage than XGBoost. This model uses leaf-centered tree segmentation.

4. Experiments and Results

Spaceship Titanic

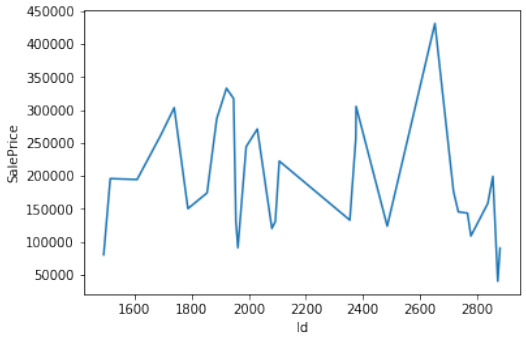
For this data, six models were used. First of all, KNN, Gaussian Naive Bayes, and logistic regression showed not very high accuracy. Therefore, we decided to use the ensemble method to increase accuracy. First, among the ensemble methods, Random Forest, which corresponds to the bagging technique, was used to achieve a high accuracy of 80.00%. The XGBoost and LightGBM: 5-folds models corresponding to the boosting technique were used, and the LightGBM model showed a stable 80.00% accuracy. Therefore, the results created based on LightGBM were stored in the data frame. And we visualized the results using the count plot of the Seaborn package.



Description: As you can see in the count plot, there are more customers who have transported than those who have not. And if you calculate the percentage, you get about 62.17% of the customers transport.

House Prices

For this data, four models were used. First of all, the Linear Regression and Lasso models showed approximately 85.00% fit. Therefore, we used LightBM and XGBoost models that use boosting techniques among ensemble methods. Both models had a fit of more than 90.00%. As a result, the results of the models using the ensemble method were blended by 50.00% each and stored in a data frame.



Description: This data contains too much data, so 30 samples were taken to create a line plot. If you look at the graph, you don't see the pattern, but you can see that most houses cost more than $100,000. And we can see that the most expensive house price among the sampled data is about $450,000.

5. Summary and Conclusions

Spaceship Titanic

When using ensemble techniques to make predictions based on Titanic data, the accuracy was usually about 80.00%. Based on these results, the analysis shows that more than 50.00% of customers have been transported as alternative dimensions.

House Prices

The housing price data also showed the highest fit of approximately 90.00% when predicted through models corresponding to ensemble techniques. In particular, this data is so large that we decided to analyze it through sampled data rather than visualizing the overall results. The results showed that although it was difficult to find a pattern, it usually formed a price range of more than $100,000.

Acknowledgments

References