EXHIBIT-ART

A SOLUTION TO SHIPPING COMPANY

BY: DEEPAK JOSHI

Table Of Content

AIM

- To predict the cost required to ship the sculptures to customers based on the information provided in the dataset.
- Perform EDA to provide the Insights not only to improve the Predication quality of model but to enhance the profit of the company.

How Can This Help!!

Retention Rate

They can prioritize transports based on how reputated the artist is and how big the client is, because they don't want to mess with good paying clients.This will help company to increase their Retention Rate of customers.

Delivery Cost

They can do batch transports of the sculptures which are intended to go at same nearby location which will Reduce Delivery cost.

Business Growth

The company can make profit and estimate the profit from this by asking the relevant amount from the client which creates sustainable growth of business without over charging or facing any losses.

Target Identification

The company can also target the specific regions accordingly to demand (i.e Opening bigger warehouses in more demanding areas instead of random areas or nearby multiple normal demanding areas) and many more.

Introduction

- Delivery Is More Important Now Than Ever.
- Covid Restriction thought the world has contributed in the rise of the Delivery Services.
- Now with Increase demands market is shifting towards the Competitive Pricing.
- These companies have huge data and large scope of growth.
- Now this company wants to compete to stay in market and this project will help them .

ABOUT DATASET

Columns	Description
Customer Id	Represents the unique identification number of the customers
Artist Name	Represents the name of the artist
Artist Reputation	Represents the reputation of an artist in the market
Height	Represents the height of the sculpture
Width	Represents the width of the sculpture
Weight	Represents the weight of the sculpture

Columns	Description
Material	Represents the material that the sculpture is made of
Price Of Sculpture	Represents the price of the sculpture
Base Shipping Price	Represents the base price for shipping a sculpture
International	Represents whether the shipping is international
Express Shipment	Represents whether the shipping was in the express (fast) mode
Installation Included	Represents whether the order had installation included in the purchase

Columns	Description
Transport	Represents the mode of transport of the order
Fragile	Represents whether the order is fragile
Customer Information	Represents details about a customer
Remote Location	Represents whether the customer resides in a remote location(i.e. relatively difficult to reach location)
Scheduled Date	Represents the date when the order was placed
Delivery Date	Represents the date of delivery of the order

Methodology

- Data Collection
- EDA
- Preparing Data
- Prediction

Methodology

DATA COLLECTION

- Data was a part of competition which was held online by Hacker Rank , 3-4 years ago.
- Data was in multiple csv files.
- I was downloaded and used in the project after combining.

Methodology

EDA

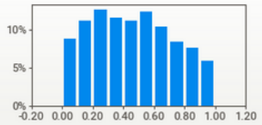
- EDA was done to find the Quality of data recieved.
- In EDA we analysed that there are few missing values and few NA values.
- During this process we also encounterd few outliers,but we cannot remove the outliers as they are meaningful to the data(i.e. they seems to be outliers but they are not).
- Data Insights were extracted from the data using various plots and graphs.
- EDA is also used to find the best data and features to feed into the Machine Learning Model.
- Also used SweetViz of Visualisation.

Methodology

EDA

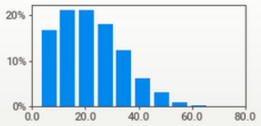
3 **Artist_Reputation**

VALUES:	5,750 (88%)	MAX	1.00	RANGE	1.00
MISSING:	750 (12%)	95%	0.91	IQR	0.440
		Q3	0.68	STD	0.266
DISTINCT:	101 (2%)	AVG	0.46	VAR	0.071
		MEDIAN	0.45		
ZEROES:	22 (<1%)	Q1	0.24	KURT.	-1.03
		5%	0.06	SKEW	0.141
		MIN	0.00	SUM	2,656



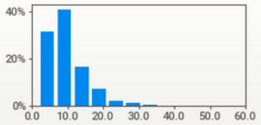
4 **Height**

VALUES:	6,125 (94%)	MAX	73.0	RANGE	70.0
MISSING:	375 (6%)	95%	44.0	IQR	18.0
		Q3	30.0	STD	12.0
DISTINCT:	65 (1%)	AVG	21.8	VAR	143
		MEDIAN	20.0		
ZEROES:	---	Q1	12.0	KURT.	-0.005
		5%	5.0	SKEW	0.594
		MIN	3.0	SUM	133k



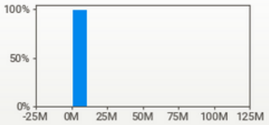
5 **Width**

VALUES:	5,916 (91%)	MAX	50.0	RANGE	48.0
MISSING:	584 (9%)	95%	21.0	IQR	6.00
		Q3	12.0	STD	5.42
DISTINCT:	40 (<1%)	AVG	9.6	VAR	29.3
		MEDIAN	8.0		
ZEROES:	---	Q1	6.0	KURT.	3.36
		5%	4.0	SKEW	1.55
		MIN	2.0	SUM	56,898



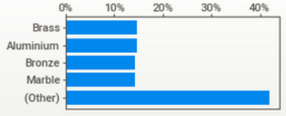
6 **Weight**

VALUES:	5,913 (91%)	MAX	117.9M	RANGE	117.9M
MISSING:	587 (9%)	95%	1.4M	IQR	35,953
		Q3	0.0M	STD	2.7M
DISTINCT:	4,410 (68%)	AVG	0.4M	VAR	7.2T
		MEDIAN	0.0M		
ZEROES:	---	Q1	0.0M	KURT.	732
		5%	0.0M	SKEW	21.6
		MIN	0.0M	SUM	2.4B



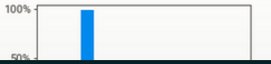
7 **Material**

VALUES:	5,736 (88%)	
MISSING:	764 (12%)	
DISTINCT:	7 (<1%)	



8 **Price_Of_Sculpture**

VALUES:	6,500 (100%)	MAX	382k	RANGE	382k
MISSING:	---	95%	4k	IQR	84.2
		Q3	0k	STD	8,820
DISTINCT:	3,424 (53%)	AVG	1k	VAR	77.8M

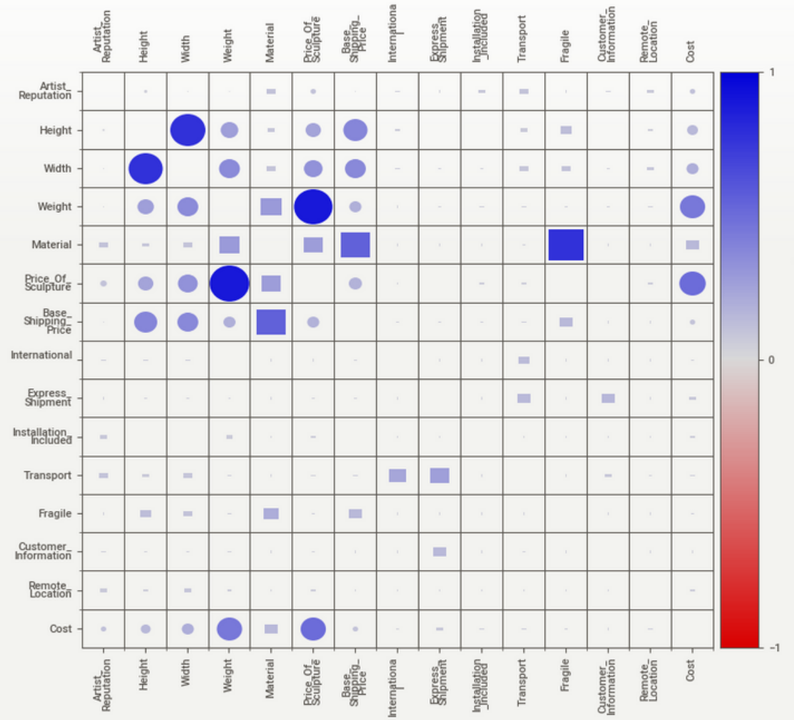


Associations

[Only including dataset "DataFrame"]

▪ **Squares** are categorical associations (uncertainty coefficient & correlation ratio) from 0 to 1. The uncertainty coefficient is **assymmetrical**, (i.e. ROW LABEL values indicate how much they PROVIDE INFORMATION to each LABEL at the TOP).

• **Circles** are the symmetrical numerical correlations (Pearson's) from -1 to 1. The **trivial diagonal** is intentionally left blank for clarity.



Methodology

EDA

	Artist_Reputation	Height	Width	Weight	Price_Of_Sculpture	Base_Shipping_Price	Cost
count	5750.000000	6125.000000	5916.000000	5.913000e+03	6500.00000	6500.000000	6.500000e+03
mean	0.461850	21.766204	9.617647	4.006948e+05	1192.42009	37.407174	1.713920e+04
std	0.265781	11.968192	5.417000	2.678081e+06	8819.61675	26.873519	2.406579e+05
min	0.000000	3.000000	2.000000	3.000000e+00	3.00000	10.000000	-8.801727e+05
25%	0.240000	12.000000	6.000000	5.030000e+02	5.23000	16.700000	1.884400e+02
50%	0.450000	20.000000	8.000000	3.102000e+03	8.02500	23.505000	3.820650e+02
75%	0.680000	30.000000	12.000000	3.645600e+04	89.47000	57.905000	1.156115e+03
max	1.000000	73.000000	50.000000	1.179279e+08	382385.67000	99.980000	1.114343e+07

Methodology

EDA

Columns with missing values

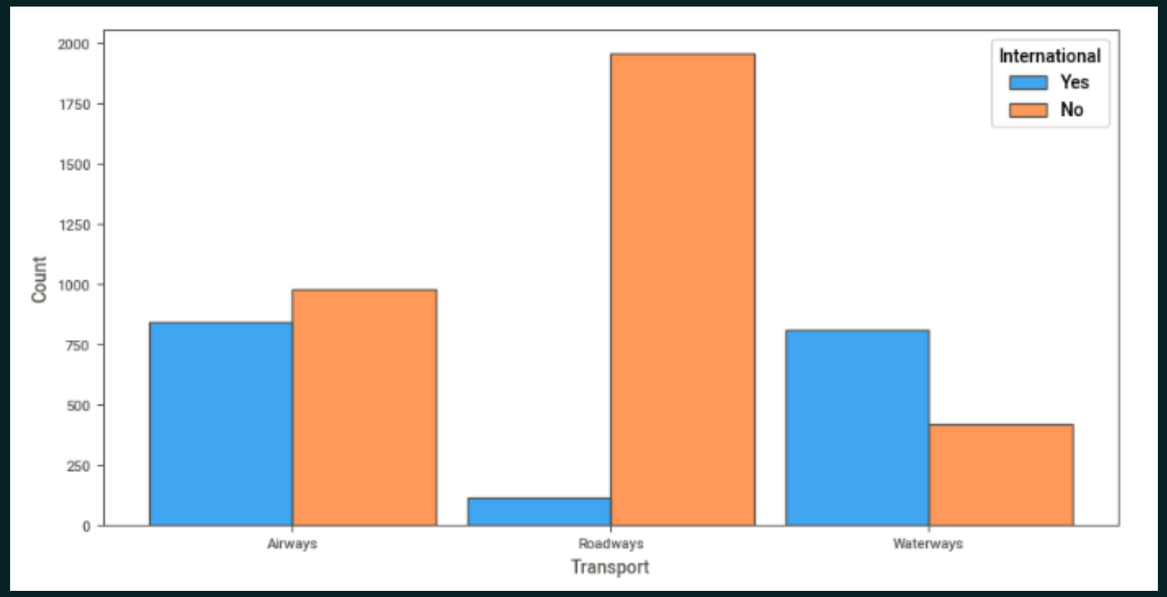
1. Artist reputation
2. Height
3. Width
4. Weight
5. Transport
6. Material
7. Remote Location

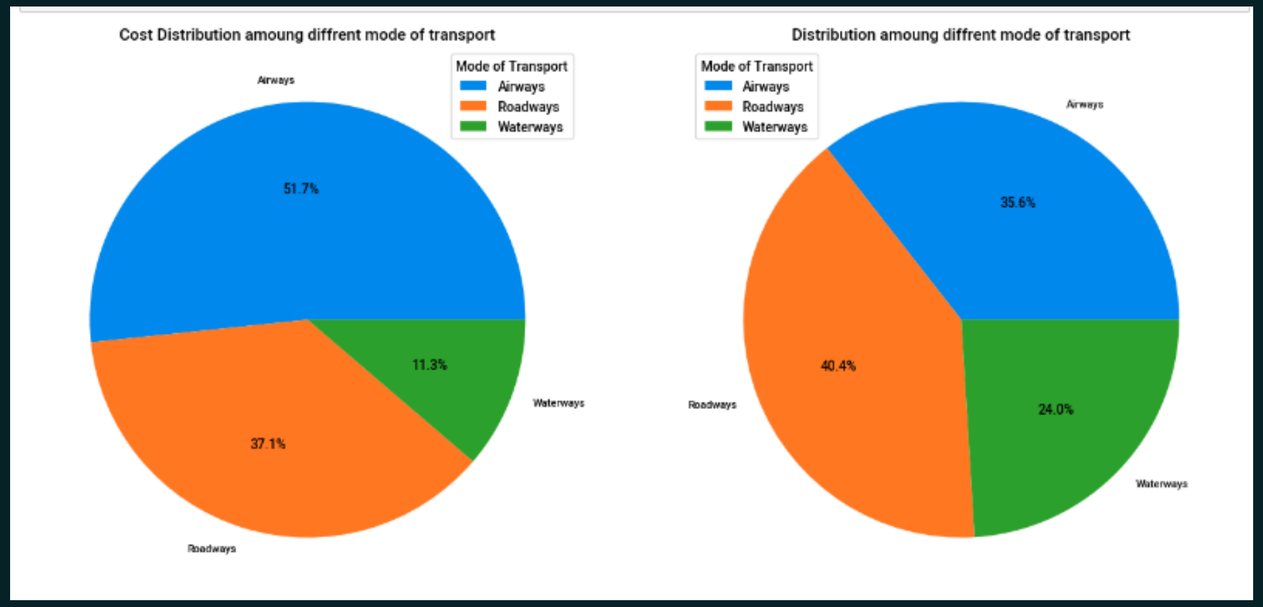
Check for categorical columns

1. Customer Id
2. Artist Name
3. Material
4. International
5. Express Shipment
6. Installation included
7. Transport
8. Fragile
9. Customer Information
10. Remote Information
11. Scheduled Date
12. Delivery Date
13. Customer Location

Methodology

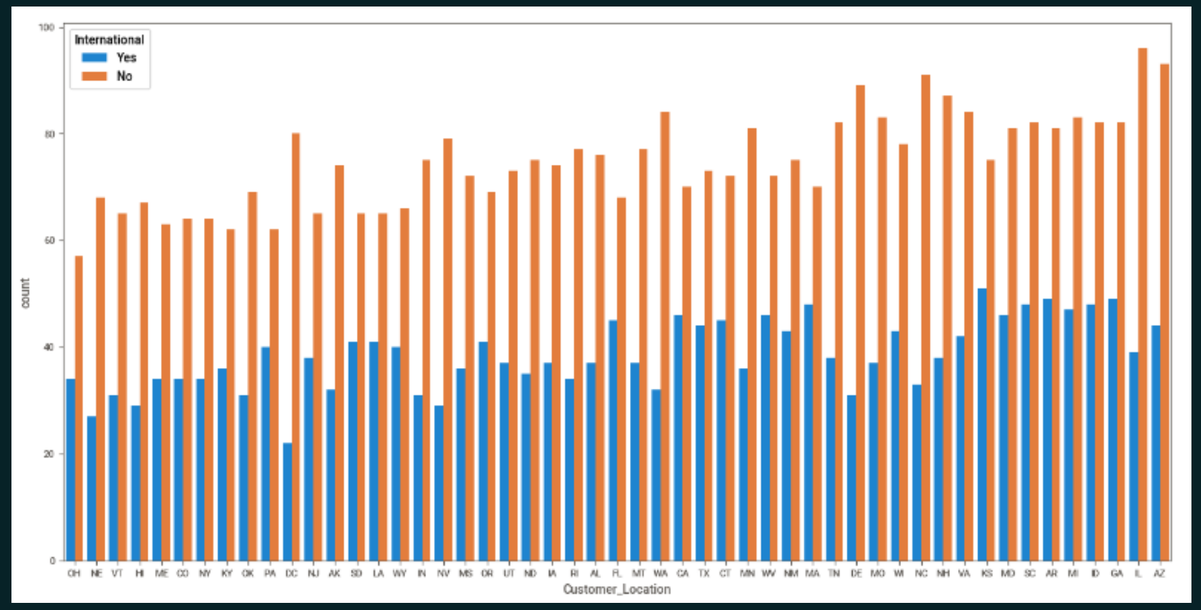
EDA

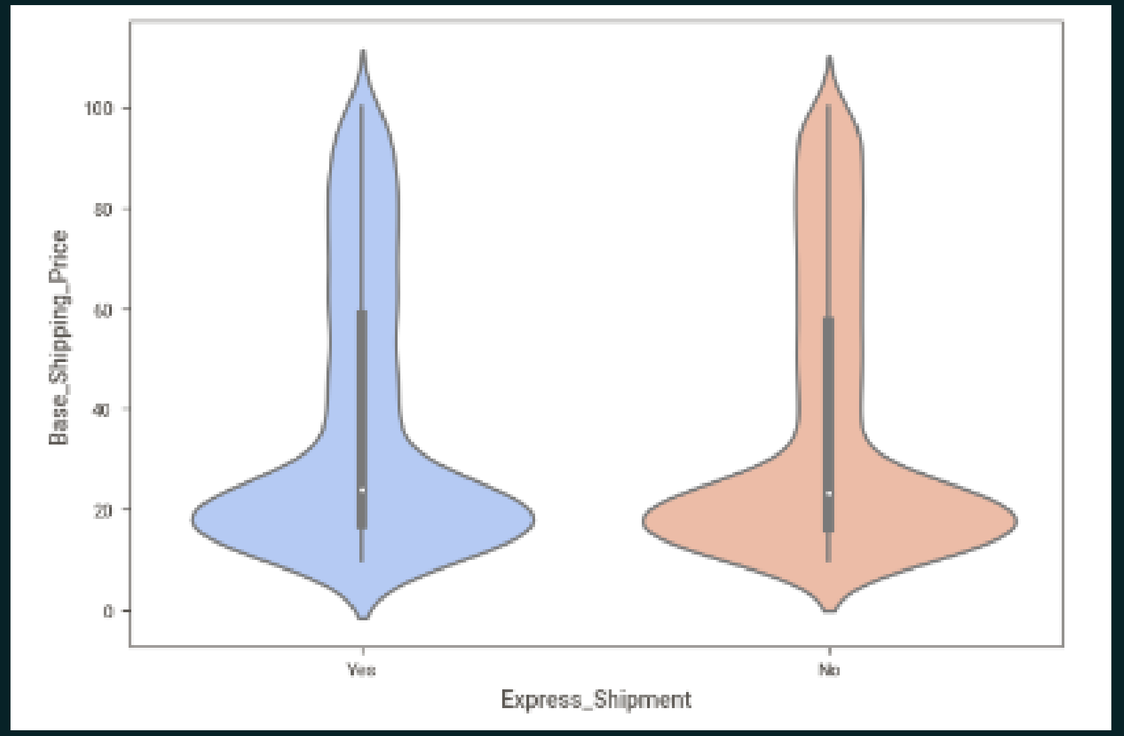




Methodology

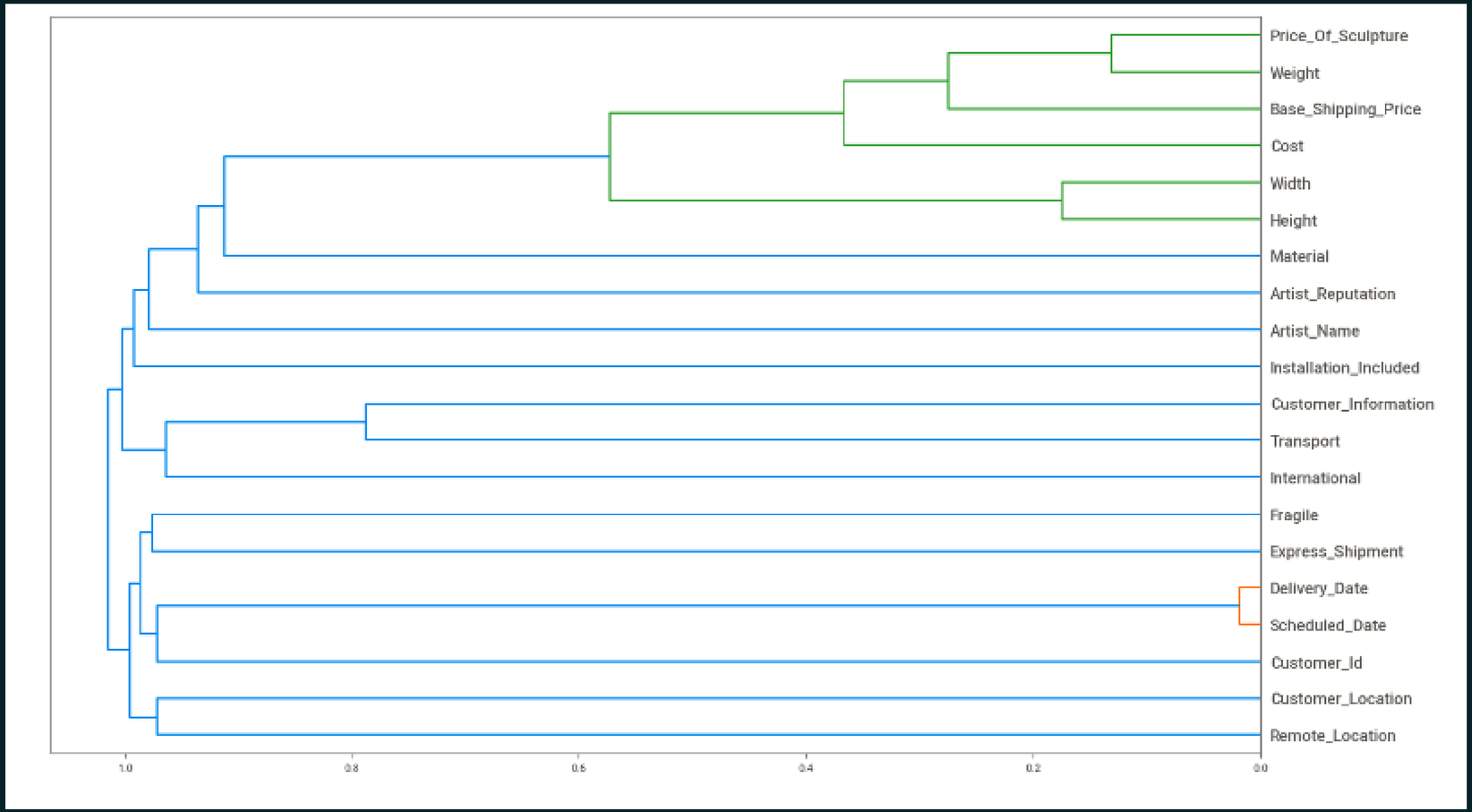
EDA





Methodology

EDA



Data Preparation

- Scaling of Data: as few alogithms work on scaled data.
- Imputing missing values & removing unnecessary features from ML model point of view.
- Converting categorical data into label of numerical data to feed it into ML models.
- Spliting Data into Train set and Test Set for validation.

Prediction

- Supervised Machine Learning is used.
- Models used:

1. Linear Regression
2. Decision Tree Regressor
3. RandomForest Regressor
4. K-Neighbour Regressor
5. Gradient Boosting Regressor
6. XGBoost Regressor
7. AdaBoost Regressor

- Also performed GridSearchCV from hyperparameter tuning.

Prediction

Result with Default Parameters

	Metric	Lr	Dtree	Forest	Knn	GBR	Xboost	AdaBoost
0	rmse	1.636932	0.0	0.133646	1.384765	0.298574	0.099511	0.624303
1	MedAE	1.027421	0.0	0.050317	0.516415	0.131106	0.045797	0.503775
2	MAE	1.207314	0.0	0.081032	0.914407	0.197807	0.067280	0.529706
3	R-squared	-13510.422024	1.0	0.993096	-0.466383	0.964289	0.996267	0.822285

Train

Test

	Metric	Lr	Dtree	Forest	Knn	GBR	Xboost	AdaBoost
0	rmse	1.689103e+00	0.510100	0.337503	1.445492	0.348467	0.291209	0.638850
1	MedAE	1.034295e+00	0.211759	0.143383	0.578381	0.149160	0.121616	0.502725
2	MAE	1.236444e+00	0.325022	0.221551	0.969740	0.228289	0.188469	0.537844
3	R-squared	-2.021174e+11	0.911516	0.957889	-0.627091	0.954600	0.969028	0.830002

Prediction

Result with Best Model via HyperParameter

	Metric	Dtree	Forest	Knn	GBR	Xboost	AdaBoost
0	rmse	1.021254	0.361125	0.744585	0.114937	0.176043	0.577062
1	MedAE	0.398436	0.184616	0.331147	0.058729	0.085726	0.322195
2	MAE	0.654510	0.253781	0.489173	0.080395	0.122139	0.422413
3	R-squared	0.362799	0.946768	0.694151	0.995008	0.988183	0.856202

Train

Test

	Metric	Dtree	Forest	Knn	GBR	Xboost	AdaBoost
0	rmse	1.055696	0.404247	1.017206	0.275991	0.286961	0.578568
1	MedAE	0.405045	0.204633	0.419798	0.112037	0.122579	0.333632
2	MAE	0.679008	0.285709	0.662366	0.176780	0.188481	0.428971
3	R-squared	0.321072	0.937921	0.423256	0.972589	0.969926	0.867631

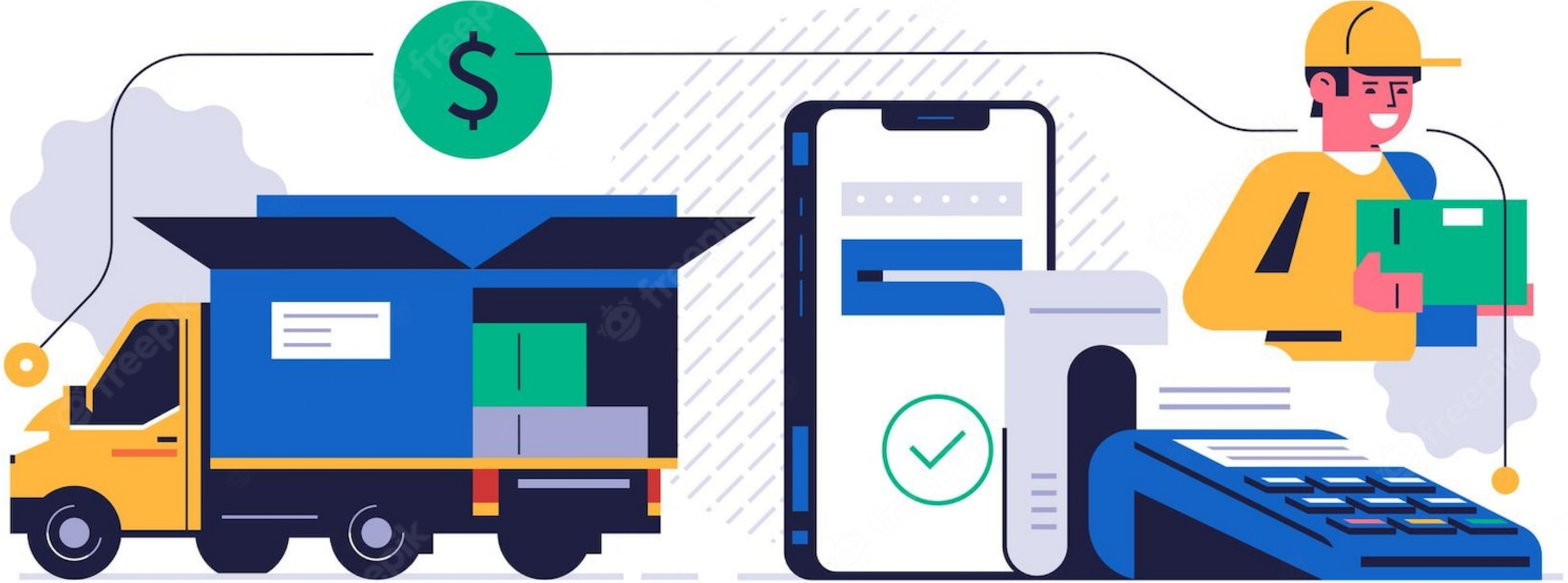
Best Model

```
%%time
# create parameters dict for tunning
XGB_para_grid = {"learning_rate"    : [0.05, 0.10, 0.15, 0.20, 0.25, 0.30 ] ,
 "max_depth"        : [ 3, 4, 5, 6, 8, 10, 12, 15],
 "min_child_weight" : [ 1, 3, 5, 7 ],
 "gamma"            : [ 0.0, 0.1, 0.2 , 0.3, 0.4 ],
 "colsample_bytree" : [ 0.3, 0.4, 0.5 , 0.7 ] }

# passing data for hyper parameter tunning with Randomized search cv
random_Search(xgbr,X_train,Y_train,XGB_para_grid)
```

```
 Best parameters:  {'min_child_weight': 7, 'max_depth': 8, 'learning_rate': 0.2
5, 'gamma': 0.4, 'colsample_bytree': 0.7}
CPU times: total: 3.84 s
Wall time: 5.98 s
```

XtremeGradient Boost Regressor



Conclusion

- We did training and prediction on various models and selected XtremeGradient Boosting Model as final model as it performed well compard to other models with an **accuracy of 98.8% on train data and around 96.9% on test data**.

- Performed EDA, preprocessing, build different models, visualized feature importance, did hyperparameter tunning of each model and did prediction.

APPENDIX

GitHub repository url: https://github.com/deepakjoshi2k/Machine-Learning-Exhibit-Art-Shipping-

Direct Download HTML: https://drive.google.com/uc?export=download&id=1NsJ2VGxQ8SAMCNVCiFGjYt-eZMRKsoBy

THANK YOU