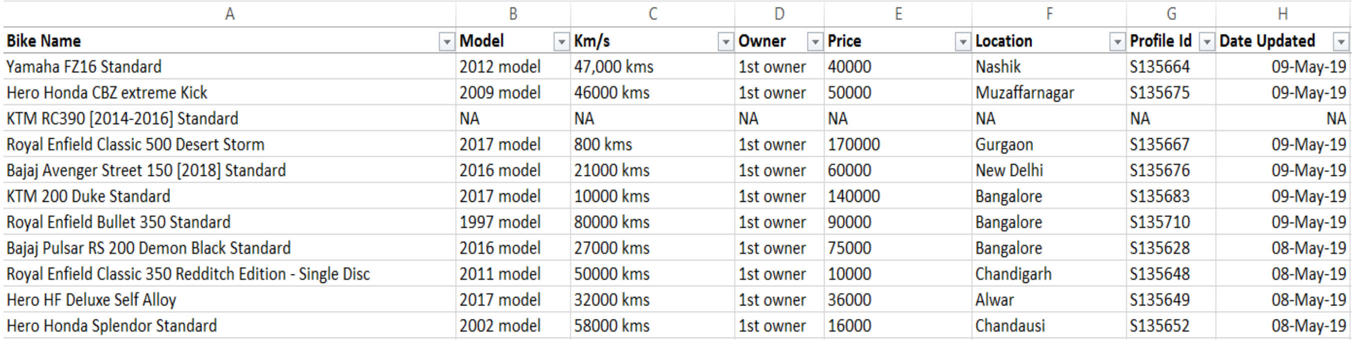
Problem Statement:

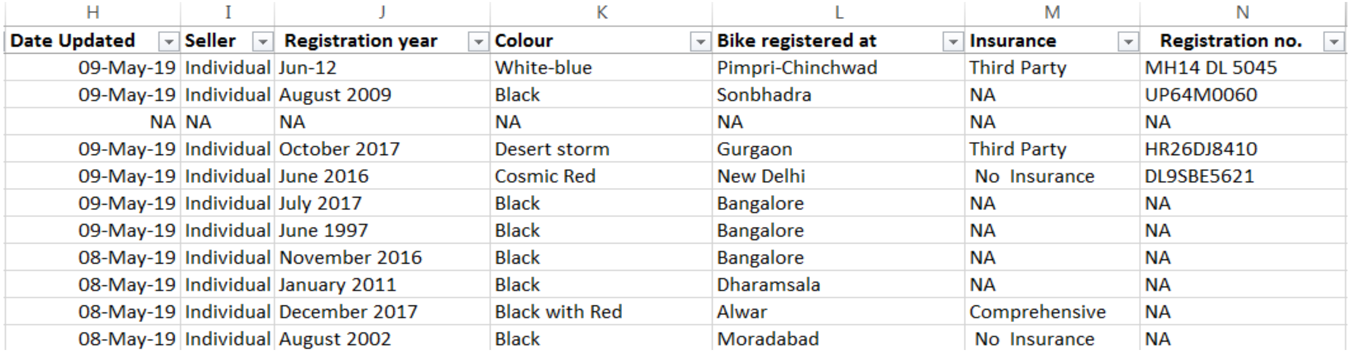
Various companies or other sellers sell used bikes providing best resale value for the product considering various parameters. These companies' tries to set this resale price in a descent segment, so customers are attracted to these prices.
From given Data, extract insights through various parameters of the original price range of the products and predict a reasonable Resale Price of these Bikes (as regression) and predict Price Range for the same (as classification).

Dataset Description:

- It contains Bike name with its model number, Distance Covered till data updated along with the Price.
- Column name 'Owner' represent the position of owner ('2nd owner': purchased from another customer) along with the location and seller with the time when this data is lastly updated.
- It also contains Profile id, Registration no., color of the bike and the registration year of the bike.



A	B	C	D	E	F	G	H
Bike Name	Model	Km/s	Owner	Price	Location	Profile Id	Date Updated
Yamaha FZ16 Standard	2012 model	47,000 kms	1st owner	40000	Nashik	S135664	09-May-19
Hero Honda CBZ extreme Kick	2009 model	46000 kms	1st owner	50000	Muzaffarnagar	S135675	09-May-19
KTM RC390 [2014-2016] Standard	NA	NA	NA	NA	NA	NA	NA
Royal Enfield Classic 500 Desert Storm	2017 model	800 kms	1st owner	170000	Gurgaon	S135667	09-May-19
Bajaj Avenger Street 150 [2018] Standard	2016 model	21000 kms	1st owner	60000	New Delhi	S135676	09-May-19
KTM 200 Duke Standard	2017 model	10000 kms	1st owner	140000	Bangalore	S135683	09-May-19
Royal Enfield Bullet 350 Standard	1997 model	80000 kms	1st owner	90000	Bangalore	S135710	09-May-19
Bajaj Pulsar RS 200 Demon Black Standard	2016 model	27000 kms	1st owner	75000	Bangalore	S135628	08-May-19
Royal Enfield Classic 350 Redditch Edition - Single Disc	2011 model	50000 kms	1st owner	10000	Chandigarh	S135648	08-May-19
Hero HF Deluxe Self Alloy	2017 model	32000 kms	1st owner	36000	Alwar	S135649	08-May-19
Hero Honda Splendor Standard	2002 model	58000 kms	1st owner	16000	Chandausi	S135652	08-May-19



H	I	J	K	L	M	N
Date Updated	Seller	Registration year	Colour	Bike registered at	Insurance	Registration no.
09-May-19	Individual	Jun-12	White-blue	Pimpri-Chinchwad	Third Party	MH14 DL 5045
09-May-19	Individual	August 2009	Black	Sonbhadra	NA	UP64M0060
NA	NA	NA	NA	NA	NA	NA
09-May-19	Individual	October 2017	Desert storm	Gurgaon	Third Party	HR26DJ8410
09-May-19	Individual	June 2016	Cosmic Red	New Delhi	No Insurance	DL9SBE5621
09-May-19	Individual	July 2017	Black	Bangalore	NA	NA
09-May-19	Individual	June 1997	Black	Bangalore	NA	NA
08-May-19	Individual	November 2016	Black	Bangalore	NA	NA
08-May-19	Individual	January 2011	Black	Dharamsala	NA	NA
08-May-19	Individual	December 2017	Black with Red	Alwar	Comprehensive	NA
08-May-19	Individual	August 2002	Black	Moradabad	No Insurance	NA

Tools Used:

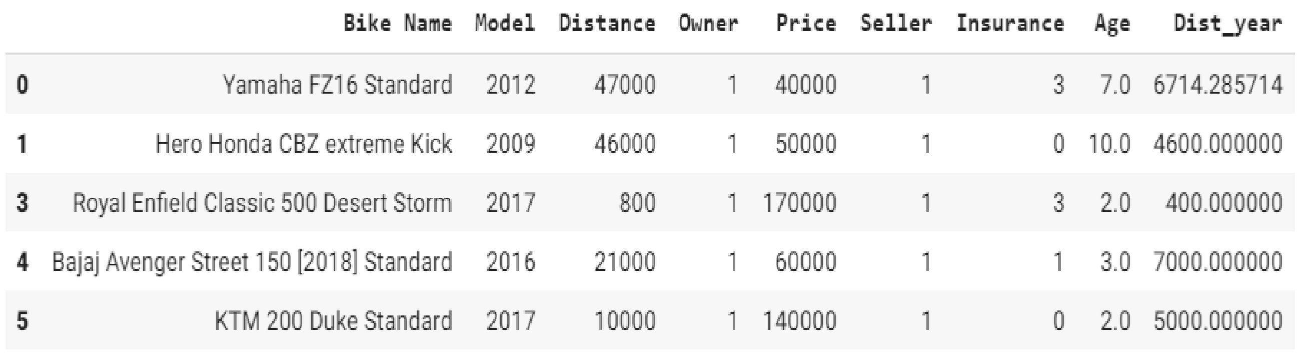
Coding Language: Python 3.0
Libraries: Pandas, Numpy, Matplotlib, Seaborn, sklearn, math, datetime
Platform: PyCharm
Algorithm: Linear Regression, RandomForest Regression, XGBoost Regression,

Solution:

This data contained data for many owners (2nd owner, 3rd …) but we only focused on 1st owner to predict the resale price for 2nd owner. Also some new features are calculated like Age, Variation of distance travelled with age etc. Then, this data is tested with various algorithms to predict resale price of bike. This problem is solved with two ways: Firstly as Regression problem and secondly as Classification problem.

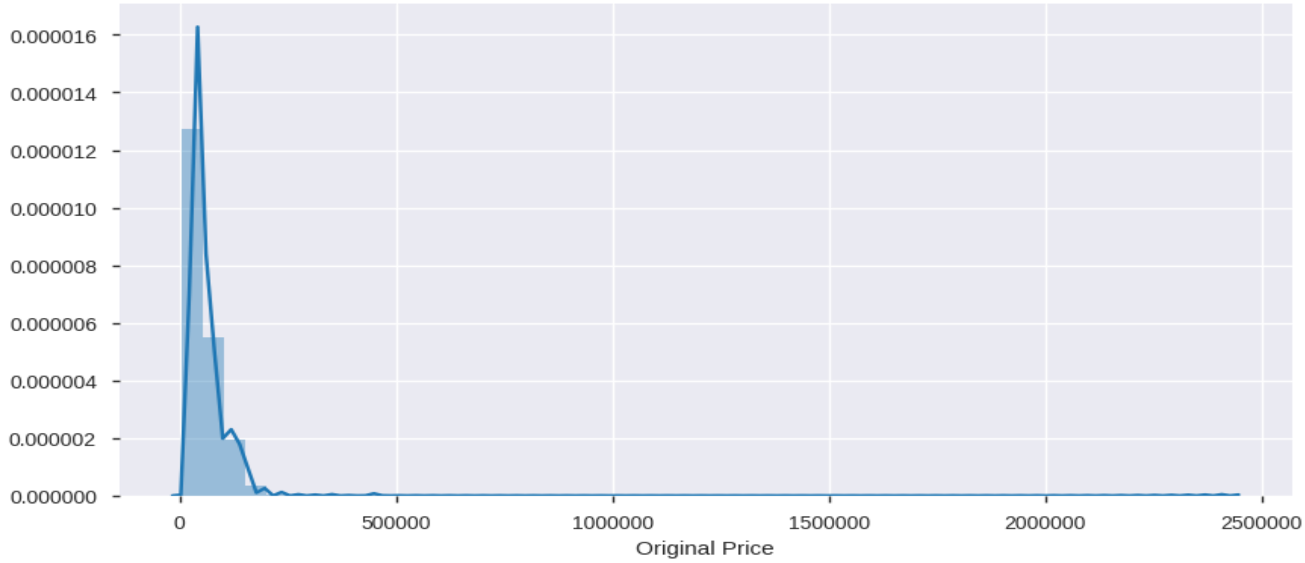
Steps:

- Importing required Libraries.
- Dropping the unwanted columns in the data and loading the Data through pandas (as df) and also renaming some columns as per requirement.
- General overview of the data like checking for the shape of the data, df.info (), if any null value exits in the data. And some plots are plotted for data profiling.
- Checking for the model and the registration year as the same year and it is same for all data.
- Cleaning the columns Insurance and Registration year (removing ‘\t’ and white spaces in them) and replacing the Nan value with zero for this column and then manually encoding the Insurance column.
- Now, dropping the Nan values as the number of Nan values has been reduced by previous step because maximum Nan values were contained in those columns.
- Converting columns: Model and Distance to integer and undergoes through some preprocessing.
- Now, some new features are created like Age of the Bike, ‘Dist_year’ as Distance travelled per age of the bike and ‘Price_dist’ as Price value with respect to the ‘Dist_year’.
- Now converting the owner column into integer and manual encoding and same is done for
- Dropping some unrequired column: ‘Registration year’, ‘Bike registered at’, ‘Registration no.’.
- Now, bikes with 1st owner are saved in new dataframe (df1), 2nd owner in df2 and 3rd owner in df3.

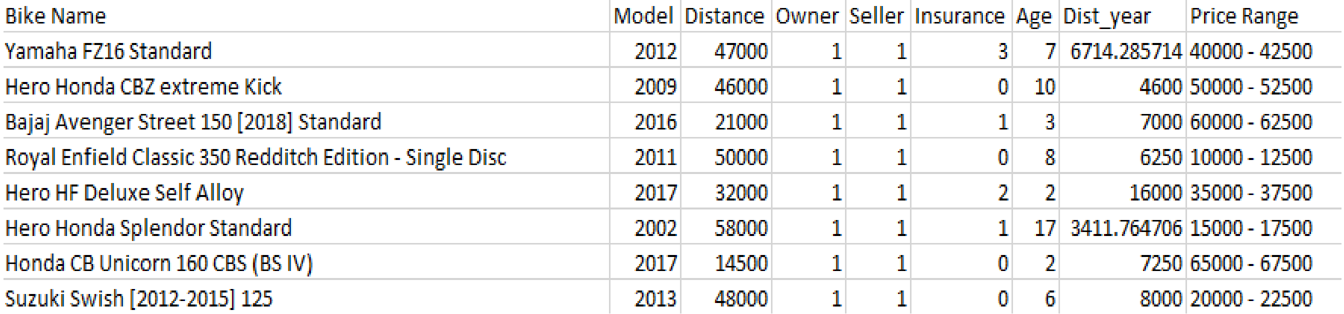


	Bike Name	Model	Distance	Owner	Price	Seller	Insurance	Age	Dist_year
0	Yamaha FZ16 Standard	2012	47000	1	40000	1	3	7.0	6714.285714
1	Hero Honda CBZ extreme Kick	2009	46000	1	50000	1	0	10.0	4600.000000
3	Royal Enfield Classic 500 Desert Storm	2017	800	1	170000	1	3	2.0	400.000000
4	Bajaj Avenger Street 150 [2018] Standard	2016	21000	1	60000	1	1	3.0	7000.000000
5	KTM 200 Duke Standard	2017	10000	1	140000	1	0	2.0	5000.000000

- Firstly it is solved as regression, now from the final dataframe outliers are removed for distance and price.
- Now, data is split into train-test set and then it fed to various models like Linear Regression, RandomForest Regressor and XGBoost Regressor.
- Then the best model is selected as per Mean Absolute Percentage Error (evaluation metrics).
- This above selection is enhanced further by doing Hyper-parameter tuning using GridSearchCV on both the models, tuning with list of parameters and then selecting the best parameter combination for the model give best evaluation score. And then Price is predicted using that model.

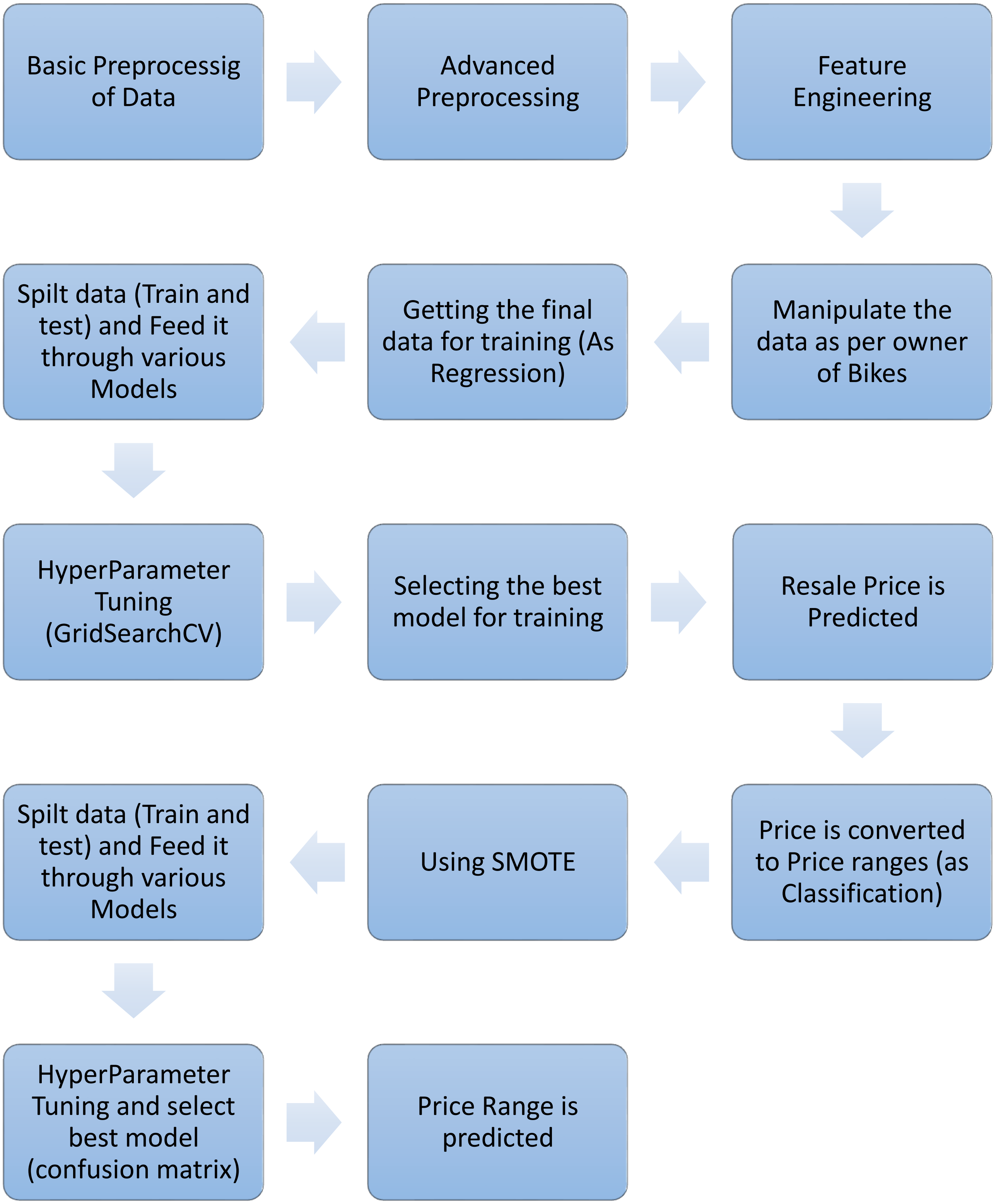


- Now, it is solved as Classification problem, a function is created which convert the Price into various Price Ranges. But this grouping is done differently for price as from 5k to 77k price, it was divided with 2.5k price gap and 77k to 200k, it was divided with 6k price gap.
- After this Price ranges are processed, it is check for imbalance and this data shows high imbalance classes so, SMOTE algorithm is used to balance these classes.
- Then data is split into train-test set and fed through various models like RandomForest Classifier and XGBoost Classifier. And then best algorithm is selected for training with confusion matrix and accuracy as evaluation metrics for the same



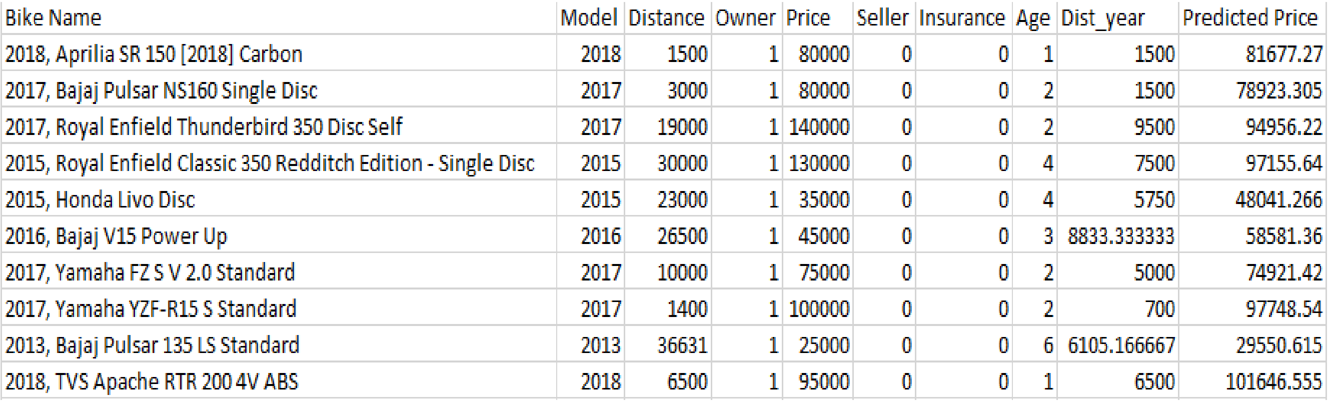
Bike Name	Model	Distance	Owner	Seller	Insurance	Age	Dist_year	Price Range
Yamaha FZ16 Standard	2012	47000	1	1	3	7	6714.285714	40000 - 42500
Hero Honda CBZ extreme Kick	2009	46000	1	1	0	10	4600	50000 - 52500
Bajaj Avenger Street 150 [2018] Standard	2016	21000	1	1	1	3	7000	60000 - 62500
Royal Enfield Classic 350 Redditch Edition - Single Disc	2011	50000	1	1	0	8	6250	10000 - 12500
Hero HF Deluxe Self Alloy	2017	32000	1	1	2	2	16000	35000 - 37500
Hero Honda Splendor Standard	2002	58000	1	1	1	17	3411.764706	15000 - 17500
Honda CB Unicorn 160 CBS (BS IV)	2017	14500	1	1	0	2	7250	65000 - 67500
Suzuki Swish [2012-2015] 125	2013	48000	1	1	0	6	8000	20000 - 22500

Flow Chart:


Basic Preprocessig of Data
Advanced Preprocessing
Feature Engineering
Manipulate the data as per owner of Bikes
Getting the final data for training (As Regression)
Spilt data (Train and test) and Feed it through various Models
HyperParameter Tuning (GridSearchCV)
Selecting the best model for training
Resale Price is Predicted
Price is converted to Price ranges (as Classification)
Using SMOTE
Spilt data (Train and test) and Feed it through various Models
HyperParameter Tuning and select best model (confusion matrix)
Price Range is predicted

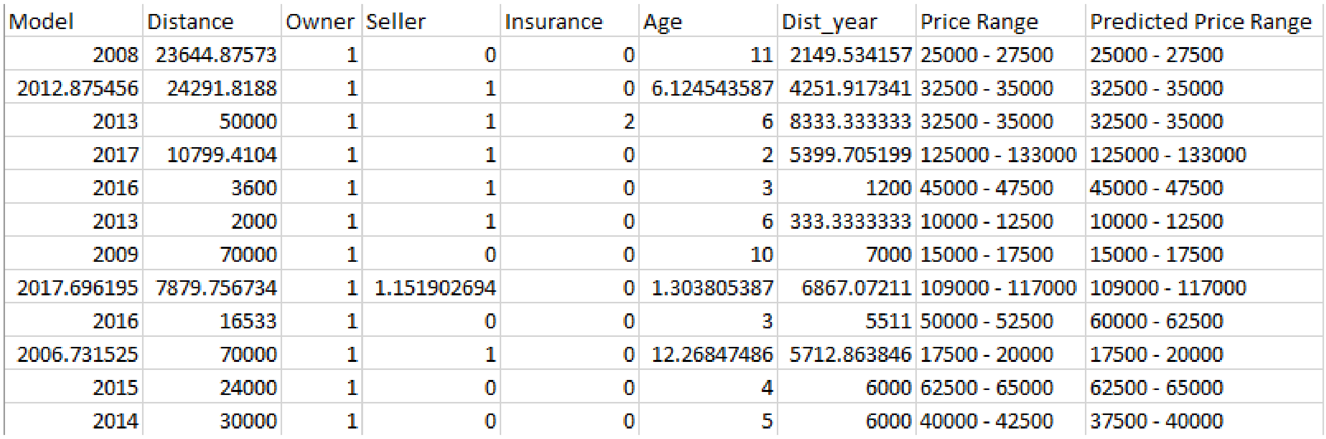
Output:

Predicted Resale Price (Regression)



Bike Name	Model	Distance	Owner	Price	Seller	Insurance	Age	Dist_year	Predicted Price
2018, Aprilia SR 150 [2018] Carbon	2018	1500	1	80000	0	0	1	1500	81677.27
2017, Bajaj Pulsar NS160 Single Disc	2017	3000	1	80000	0	0	2	1500	78923.305
2017, Royal Enfield Thunderbird 350 Disc Self	2017	19000	1	140000	0	0	2	9500	94956.22
2015, Royal Enfield Classic 350 Redditch Edition - Single Disc	2015	30000	1	130000	0	0	4	7500	97155.64
2015, Honda Livo Disc	2015	23000	1	35000	0	0	4	5750	48041.266
2016, Bajaj V15 Power Up	2016	26500	1	45000	0	0	3	8833.333333	58581.36
2017, Yamaha FZ S V 2.0 Standard	2017	10000	1	75000	0	0	2	5000	74921.42
2017, Yamaha YZF-R15 S Standard	2017	1400	1	100000	0	0	2	700	97748.54
2013, Bajaj Pulsar 135 LS Standard	2013	36631	1	25000	0	0	6	6105.166667	29550.615
2018, TVS Apache RTR 200 4V ABS	2018	6500	1	95000	0	0	1	6500	101646.555

Predicted Resale Price Range (Classification)



Model	Distance	Owner	Seller	Insurance	Age	Dist_year	Price Range	Predicted Price Range
2008	23644.87573	1	0	0	11	2149.534157	25000 - 27500	25000 - 27500
2012.875456	24291.8188	1	1	0	6.124543587	4251.917341	32500 - 35000	32500 - 35000
2013	50000	1	1	2	6	8333.333333	32500 - 35000	32500 - 35000
2017	10799.4104	1	1	0	2	5399.705199	125000 - 133000	125000 - 133000
2016	3600	1	1	0	3	1200	45000 - 47500	45000 - 47500
2013	2000	1	1	0	6	333.3333333	10000 - 12500	10000 - 12500
2009	70000	1	0	0	10	7000	15000 - 17500	15000 - 17500
2017.696195	7879.756734	1	1.151902694	0	1.303805387	6867.07211	109000 - 117000	109000 - 117000
2016	16533	1	0	0	3	5511	50000 - 52500	60000 - 62500
2006.731525	70000	1	1	0	12.26847486	5712.863846	17500 - 20000	17500 - 20000
2015	24000	1	0	0	4	6000	62500 - 65000	62500 - 65000
2014	30000	1	0	0	5	6000	40000 - 42500	37500 - 40000