SCS 3251 049: GROUP 6

BUILDING A PREDICTIVE MODEL OF USED VEHICLE PRICES USING MULTIPLE LINEAR REGRESSION

PRESENTED BY: **FILSAN MUSA, JIAYANG WANG, NIKTA Z. YUSSEFIAN, ROXY WONG, SRUTHI R. PUTHIYOPPIL**

TABLE OF CONTENT

INTRODUCTION

SOURCE: CORPORATE DATA

DATASET: USED VEHICLES

RECORDS: 17369

CATEGORICAL VARIABLES: 6

NUMERIC VARIABLES: 12

DATASET VARIABLES

Sale Price

MSRP

Contract Start Date

Accident Degree

Odometer

Model

OBJECTIVE

IDENTIFY THE MOST SIGNIFICANT PREDICTORS OF USED VEHICLES USING A MULTIPLE LINEAR REGRESSION MODEL WITH BACKWARD ELIMINATION METHOD

METHODOLODY

1

DATA EXPLORATION

Identifying potential challenges, and establishing relationships between variables

2

DATA CLEANING & TRANSFORMATION

Dropping Variables, Imputing Missing Data, One-Hot Encoding, Fixing Data Types, Generating New Variables

3

MODEL SELECTION

Identifying the best model for our dataset

4

MODEL BUILDING

Building out the model by splitting our dataset into training and test data

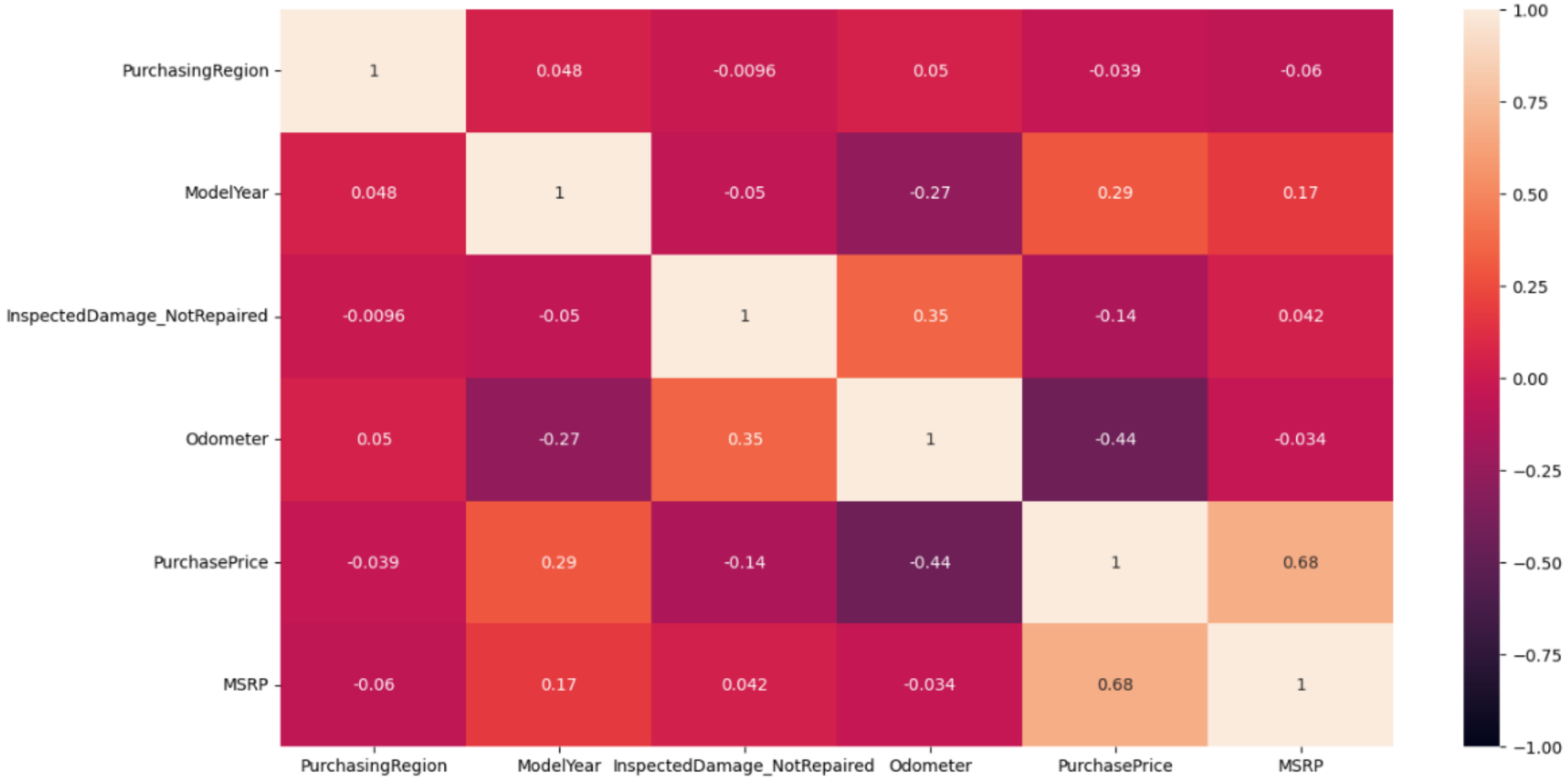
5

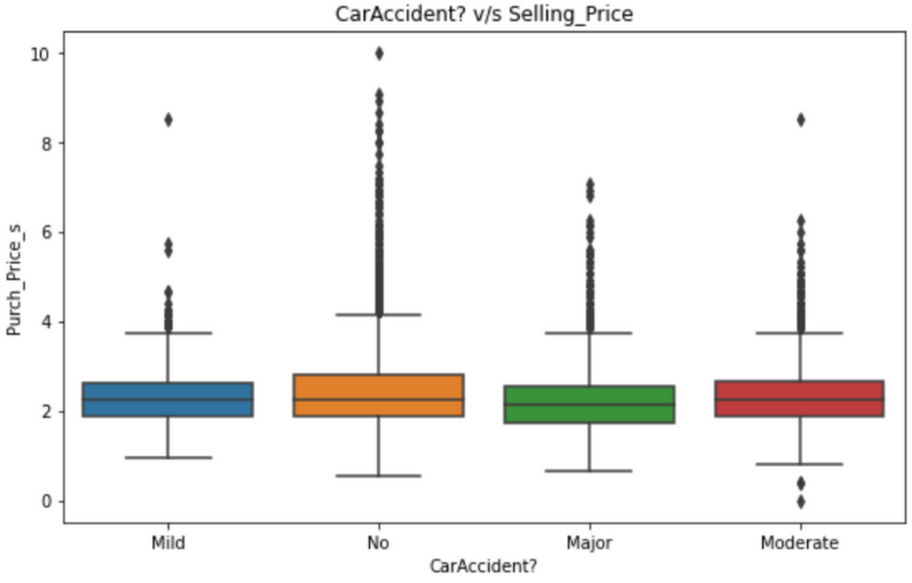
MODEL REFINEMENT

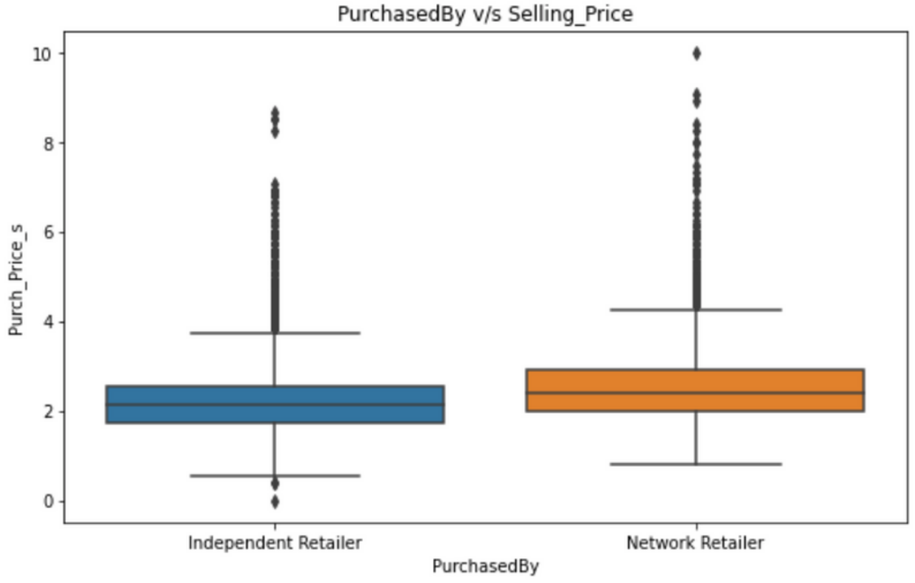
Using Backward Elimination to improve the Adj. R-Squared and other statistics

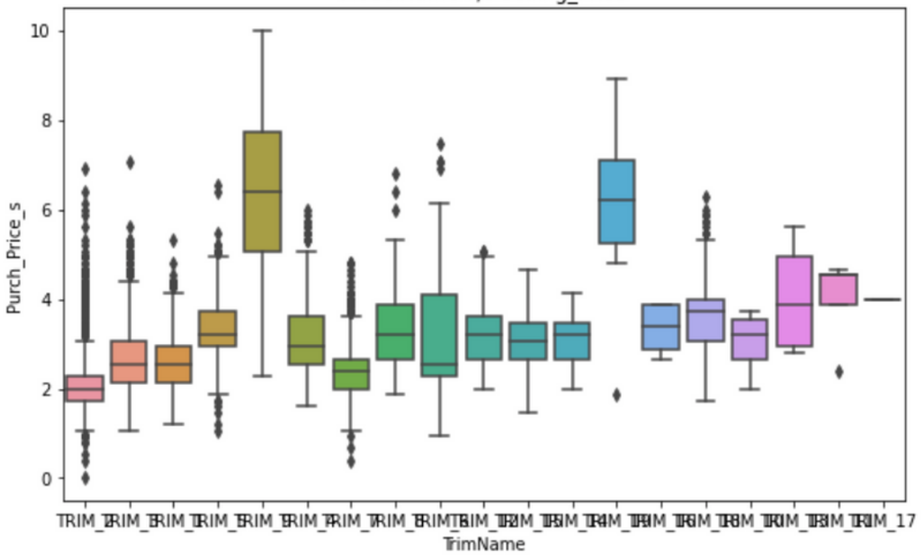
DATA PREPARATION

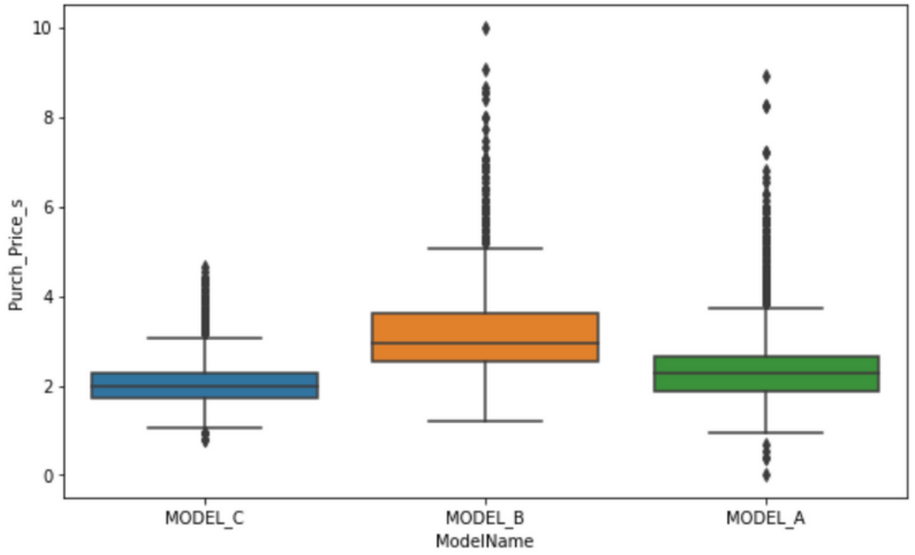
DATA EXPLORATION




CarAccident? v/s Selling_Price
Purch_Price_s
CarAccident?
Mild
No
Major
Moderate


PurchasedBy v/s Selling_Price
Purch_Price_s
PurchasedBy
Independent Retailer
Network Retailer


ModelName v/s Selling_Price
Purch_Price_s
TrimName


Purch_Price_s
MODEL_C
MODEL_B
MODEL_A
ModelName

DATA CLEANING

FIXING DATA TYPES

CHANGED THE DATA TYPE OF DATE VARIABLES INTO "DATETIME"

DROPPING VARIABLES

DROPPED 1 VARIABLE BECAUSE OF SUBSTANTIAL AMOUNTS OF MISSING VALUES

RENAMING VARIABLES

RENAMED VARIABLES TO STANDARDIZE FOR BEST PRACTICE

IMPUTING MISSING DATA

USED MEDIAN VALUE TO REPLACE MISSING VALUES

DATA TRANSFORMATION

ONE-HOT ENCODING

USED ONE-HOT ENCODING TO CONVERT CATEGORICAL VARIABLES INTO NUMERIC

ENGINEERING NEW VARIABLES

GENERATED TWO NEW VARIABLES "AGE" AND "DEPRECIATION"

SCALING VARIABLES

ADJUSTING VALUES TO FALL BETWEEN ZERO AND TEN

DATA ANALYSIS & MODELING

SUMMARY STATISTICS

OUTLIERS & QUARTILES

EXAMINING VALUES THAT DO NOT FOLLOW THE GENERAL TREND OF OUR DATA

LEVERAGE & INFLUENCE POINTS

IDENTIFYING HIGH LEVERAGE POINTS, AND HIGH INFLUENCE

STATISTICAL MOMENTS

MEAN, STD, VARIANCE, SKEWNESS, KURTOSIS

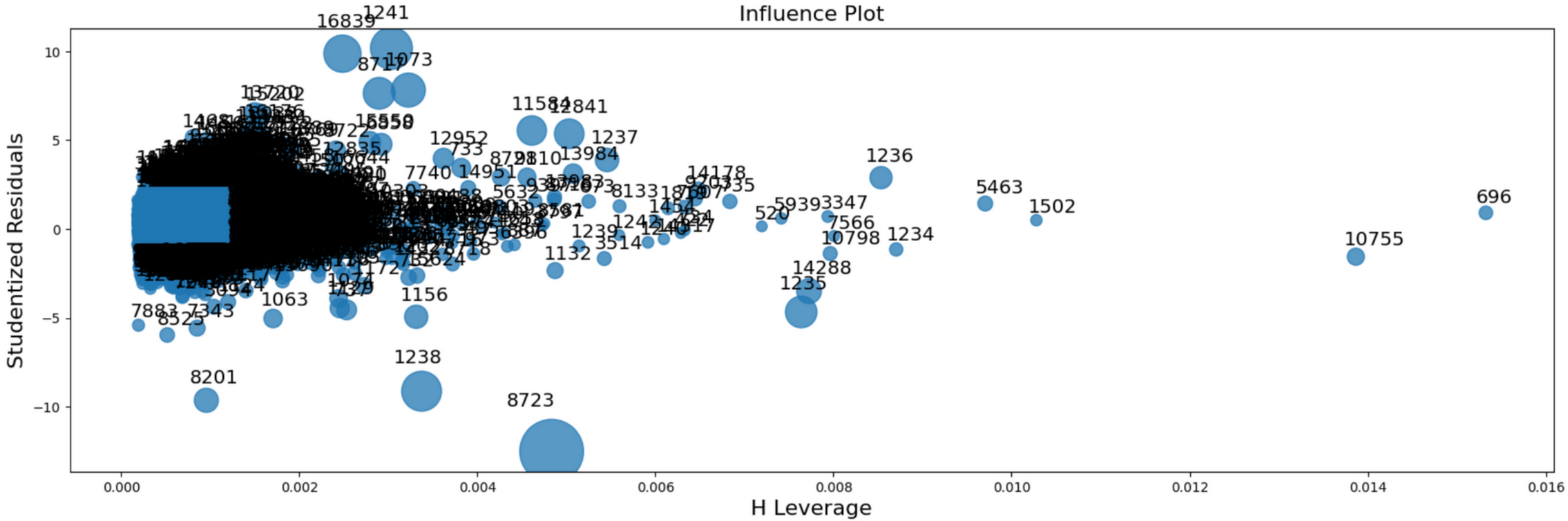
PERFORMANCE METRICS

MSE, R-SQUARED, ADJ. R-SQUARED, RMSE

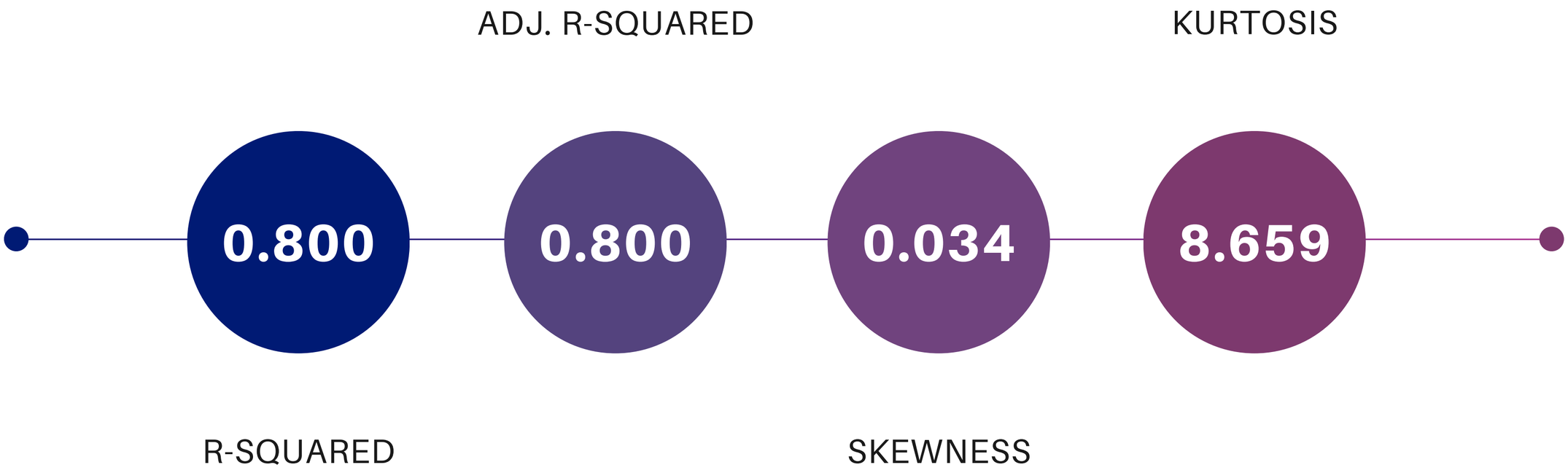
INFERENTIAL STATISTICS

P-VALUE, STANDARD ERROR

LEVERAGE & INFLUENCE POINTS



OLS REGRESSION RESULTS



OLS Regression Results

Dep. Variable:	Purch_Price_s	R-squared:	0.800
Model:	OLS	Adj. R-squared:	0.800
Method:	Least Squares	F-statistic:	3466.
Date:	Fri, 31 Mar 2023	Prob (F-statistic):	0.00
Time:	23:57:05	Log-Likelihood:	-5467.8
No. Observations:	17357	AIC:	1.098e+04
Df Residuals:	17336	BIC:	1.114e+04
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.0852	0.027	76.671	0.000	2.032	2.139
SalesChannel[T.B]	0.0635	0.009	7.346	0.000	0.047	0.080
SalesChannel[T.C]	0.0419	0.012	3.610	0.000	0.019	0.065
SalesChannel[T.D]	0.0660	0.013	4.957	0.000	0.040	0.092
SalesChannel[T.E]	0.0488	0.015	3.325	0.001	0.020	0.078
SalesChannel[T.F]	-0.0403	0.015	-2.611	0.009	-0.071	-0.010
SalesChannel[T.G]	-0.0351	0.014	-2.479	0.013	-0.063	-0.007
SalesChannel[T.H]	-0.0066	0.012	-0.569	0.569	-0.029	0.016
InspectedDamage_NotRepaired_s	-0.0240	0.004	-5.404	0.000	-0.033	-0.015
Odometer_s	-0.1183	0.002	-57.046	0.000	-0.122	-0.114
MSRP_s	0.7049	0.005	138.526	0.000	0.695	0.715
age_s	-0.2206	0.002	-93.331	0.000	-0.225	-0.216
CarAccident_Mild	0.1700	0.015	11.244	0.000	0.140	0.200
CarAccident_Moderate	0.1301	0.009	13.766	0.000	0.112	0.149
CarAccident_No	0.1861	0.008	23.500	0.000	0.171	0.202
PurchBy_Net_R	0.0720	0.009	8.174	0.000	0.055	0.089
ModelName_MODEL_B	-0.4441	0.012	-36.560	0.000	-0.468	-0.420
ModelName_MODEL_C	0.1874	0.007	26.150	0.000	0.173	0.201
PurchasingRegion_2	-0.0352	0.012	-2.832	0.005	-0.060	-0.011
PurchasingRegion_3	-0.0761	0.013	-5.789	0.000	-0.102	-0.050
PurchasingRegion_4	-0.0187	0.019	-0.964	0.335	-0.057	0.019

Omnibus:	2096.529	Durbin-Watson:	1.425
Prob(Omnibus):	0.000	Jarque-Bera (JB):	23160.221
Skew:	0.034	Prob(JB):	0.00
Kurtosis:	8.659	Cond. No.	103.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

INITIAL MODEL

$$SalePrice = 2.09 + 0.7MSRP - .22CarAge - 0.12Odometer - 0.02InspectedDamage\_NotRepaired_s$$
$$+ 0.06SalesChannel\_B + 0.04SalesChannel\_C + 0.07SalesChannel\_D + 0.05SalesChannel\_E$$
$$- 0.04SalesChannel\_F - 0.04SalesChannel\_G - 0.01SalesChannel\_H + 0.17CarAccident\_Mild$$
$$+ 0.13CarAccident\_Moderate + 0.19CarAccident\_No + 0.07PurchaseBy\_Retailer - 0.44ModelName\_ModelB$$
$$+ 0.19ModelName\_ModelB - 0.04PurchasingRegion\_2 - 0.08PurchasingRegion\_3 - 0.02PurchasingRegion\_4$$

MODEL REFINEMENT

BACKWARD ELIMINATION

IMPROVE R-SQUARED

The objective of the backward elimination step regression is to refine the model and to eliminate low impact and high leverage dependent variables

- **P-VALUES > 0.05**

  REMOVE VARIABLES WITH HIGH P-VALUE

ADDITIONAL REFINEMENT

- **HIGH STANDARD DEVIATION**

  REMOVE VARIABLES WITH HIGH STANDARD DEVIATION AFTER SCALING VARIABLES

- **HIGH LEVERAGE POINTS**

  REMOVE VARIABLES WITH HIGH LEVERAGE POINTS

RESULTS & CONCLUSION

OLS Regression Results

Dep. Variable:	Purch_Price_s	R-squared:	0.801
Model:	OLS	Adj. R-squared:	0.801
Method:	Least Squares	F-statistic:	6968.
Date:	Sat, 01 Apr 2023	Prob (F-statistic):	0.00
Time:	00:19:03	Log-Likelihood:	-5265.0
No. Observations:	17345	AIC:	1.055e+04
Df Residuals:	17334	BIC:	1.064e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.0370	0.021	96.400	0.000	1.996	2.078
InspectedDamage_NotRepaired_s	-0.0233	0.004	-5.261	0.000	-0.032	-0.015
Odometer_s	-0.1186	0.002	-58.306	0.000	-0.123	-0.115
MSRP_s	0.7069	0.005	138.937	0.000	0.697	0.717
age_s	-0.2198	0.002	-98.674	0.000	-0.224	-0.215
CarAccident_Mild	0.1667	0.015	11.179	0.000	0.137	0.196
CarAccident_Moderate	0.1279	0.009	13.753	0.000	0.110	0.146
CarAccident_No	0.1815	0.008	23.393	0.000	0.166	0.197
PurchBy_Net_R	0.1102	0.005	20.769	0.000	0.100	0.121
ModelName_MODEL_B	-0.4477	0.012	-37.084	0.000	-0.471	-0.424
ModelName_MODEL_C	0.1913	0.007	27.000	0.000	0.177	0.205

Omnibus:	1158.419	Durbin-Watson:	1.382
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5330.354
Skew:	0.134	Prob(JB):	0.00
Kurtosis:	5.702	Cond. No.	74.2

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

MODEL ASSESSMENT

0.801

ADJ. R-SQUARED

MEAN OF RESIDUALS

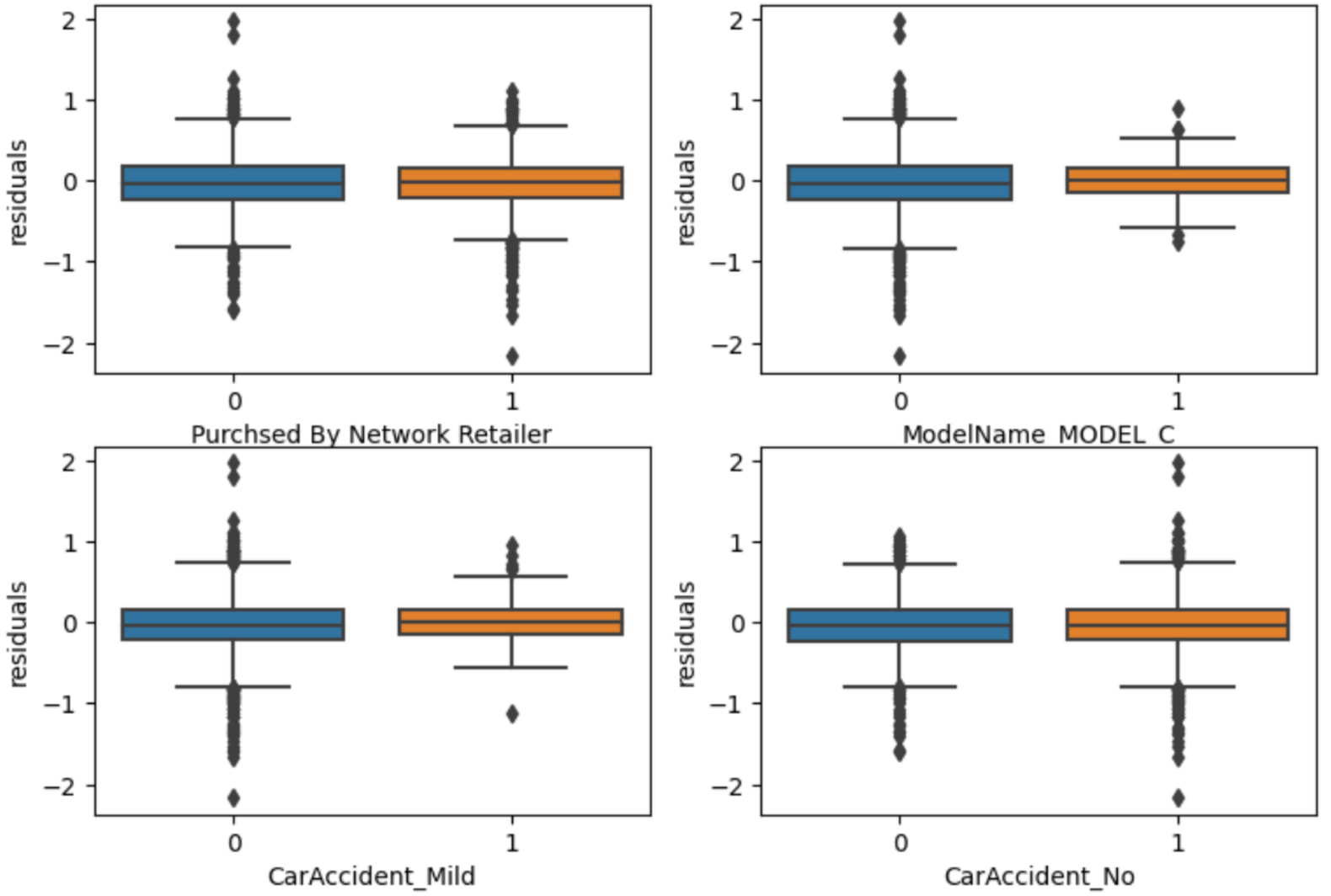
APPROX. 0

0.31

Root Mean Squared Error (RMSE)

Normalized RMSE (NRMSE)

0.17 %


residuals
2
1
0
−1
−2
0
1
Purchsed By Network Retailer
ModelName_MODEL_C
CarAccident_Mild
CarAccident_No

FINAL MODEL

$$SalePrice = 2.04 + 0.71MSRP - 0.22CarAge - 0.11Odometer - 0.45ModelName\_ModelB + 0.19ModelName\_ModelC + 0.17CarAccident\_Mild + 0.13CarAccident\_Moderate + 0.18CarAccident\_No + 0.11PurchaseBy\_Retailer$$

CONCLUSION

MODEL WAS DETERMINED A GOOD FIT AS THERE WAS RESULTING R-SQAURED > 0.80, AND MEAN OF RESIDUALS WAS CLOSE TO ZERO. IN THE FINAL VERSION OF THE MODEL THERE WERE 6 REMAINING VARIABLES: CAR AGE, ODOMETER READING, SUGGESTED RETAIL PRICE (MSRP), CAR ACCIDENT HISTORY, AND PURCHASE SOURCE.

ASSUMPTIONS & LIMITATIONS

ASSUMPTIONS

- VAGUE VARIABLE DESCRIPTIONS

  Some of the variable definition were vague, and these variables were interpreted

LIMITATIONS

- DATA DISCREPENCY

  Possible errors in some of the entries

- LIMITED APPROACH ON STEP WISE REGRESSION

  Only considered backward elimination and forward selection

- COMPROMISED NOVELTY

  To find a large and complete dataset we had to forgo the selection of a more novel datasets

END