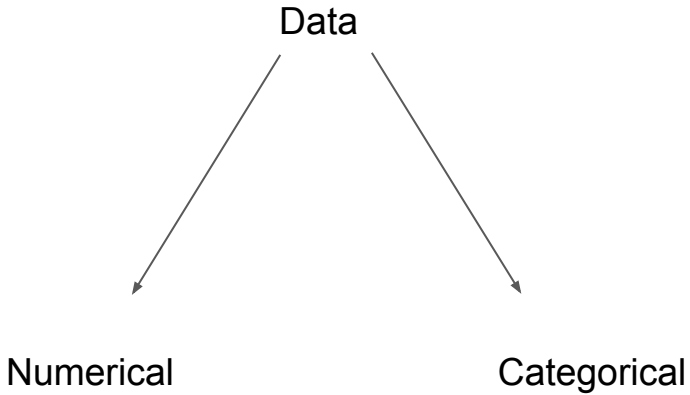
Problem Statement

What is best regression model based on the Ames Housing Dataset for predicting house prices?

How success is measured

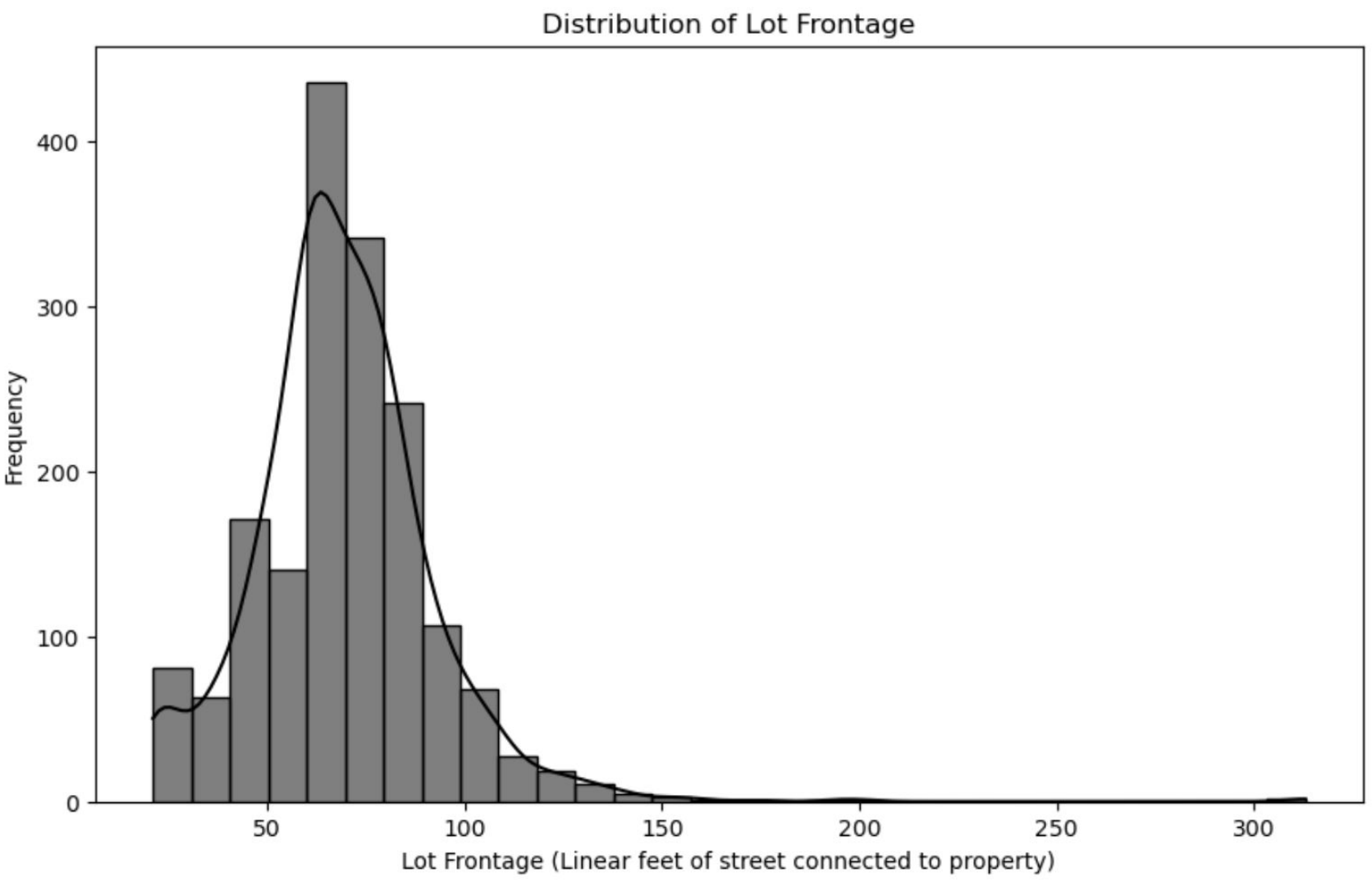
We will be using RMSE as are main score variable and linear regression as are model.

Success will be evaluated using RMSE.

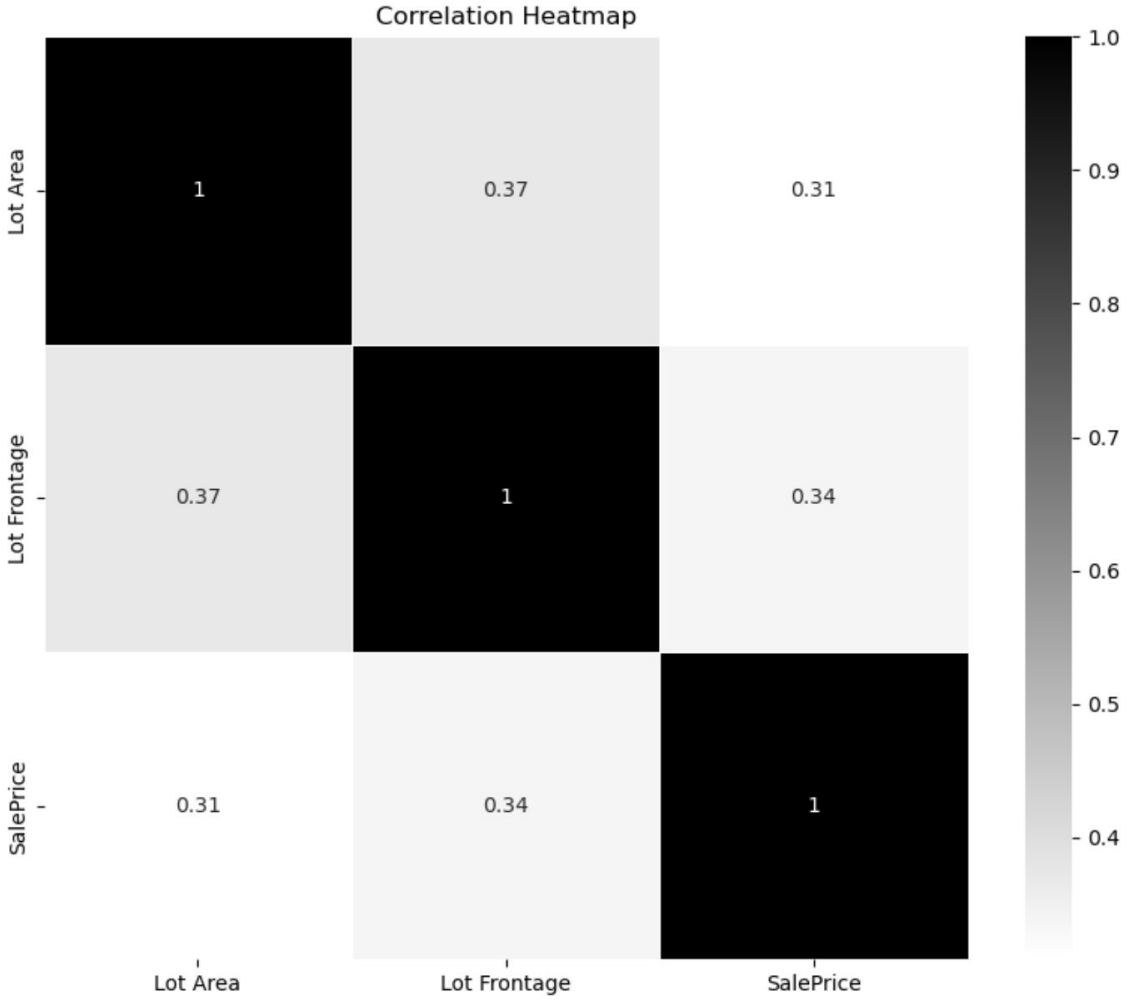

Data
Numerical
Categorical

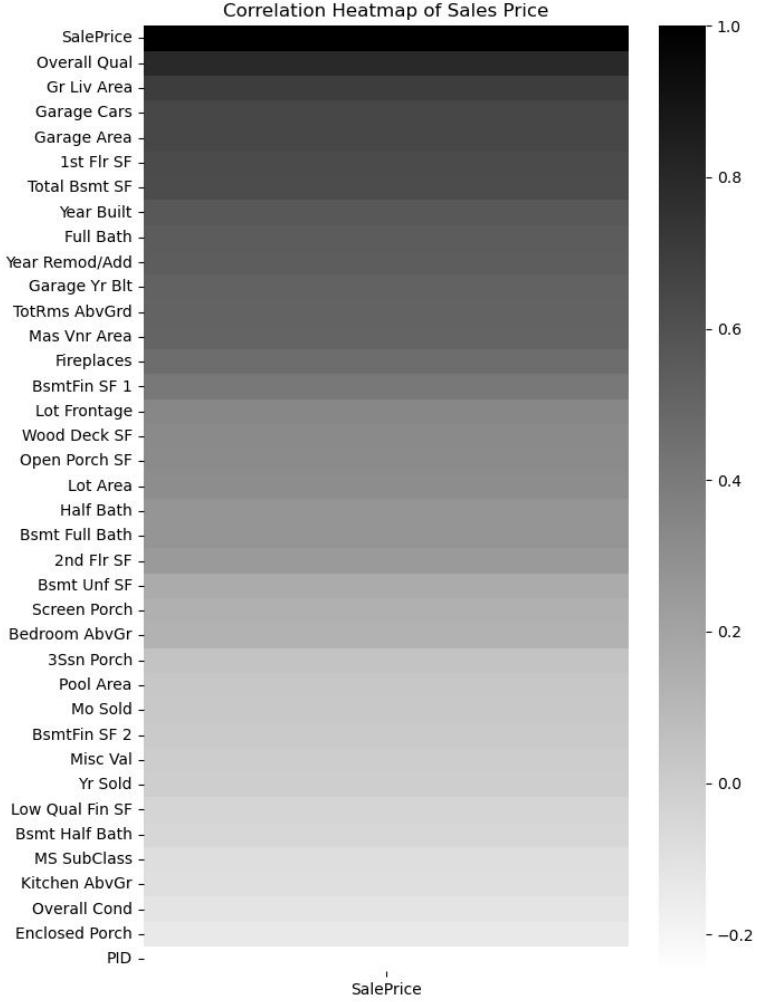
Numerical

	Missing Values	Percentage
Pool QC	2042	99.5612
Misc Feature	1986	96.8308
Alley	1911	93.1741
Fence	1651	80.4973
Fireplace Qu	1000	48.7567
Lot Frontage	330	16.0897
Garage Yr Blt	114	5.5583
Garage Cond	114	5.5583
Garage Qual	114	5.5583
Garage Finish	114	5.5583
Garage Type	113	5.5095
Bsmt Exposure	58	2.8279
BsmtFin Type 2	56	2.7304
Bsmt Cond	55	2.6816
Bsmt Qual	55	2.6816
BsmtFin Type 1	55	2.6816
Mas Vnr Area	22	1.0726
Mas Vnr Type	22	1.0726
Bsmt Half Bath	2	0.0975
Bsmt Full Bath	2	0.0975
Total Bsmt SF	1	0.0488
Bsmt Unf SF	1	0.0488
BsmtFin SF 2	1	0.0488
Garage Cars	1	0.0488
Garage Area	1	0.0488
BsmtFin SF 1	1	0.0488


Distribution of Lot Frontage
Frequency
0
100
200
300
400
50
100
150
200
250
300
Lot Frontage (Linear feet of street connected to property)

Base model


Correlation Heatmap
Lot Area
Lot Frontage
SalePrice
1
0.37
0.31
0.37
1
0.34
0.31
0.34
1
Lot Area
Lot Frontage
SalePrice
1.0
0.9
0.8
0.7
0.6
0.5
0.4


Correlation Heatmap of Sales Price
SalePrice
Overall Qual
Gr Liv Area
Garage Cars
Garage Area
1st Flr SF
Total Bsmt SF
Year Built
Full Bath
Year Remod/Add
Garage Yr Blt
TotRms AbvGrd
Mas Vnr Area
Fireplaces
BsmtFin SF 1
Lot Frontage
Wood Deck SF
Open Porch SF
Lot Area
Half Bath
Bsmt Full Bath
2nd Flr SF
Bsmt Unf SF
Screen Porch
Bedroom AbvGr
3Ssn Porch
Pool Area
Mo Sold
BsmtFin SF 2
Misc Val
Yr Sold
Low Qual Fin SF
Bsmt Half Bath
MS SubClass
Kitchen AbvGr
Overall Cond
Enclosed Porch
PID
SalePrice
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

```
('rmse = 40040.92083360502', 'r2 = 0.7387224982665793')
```

With Cross validation

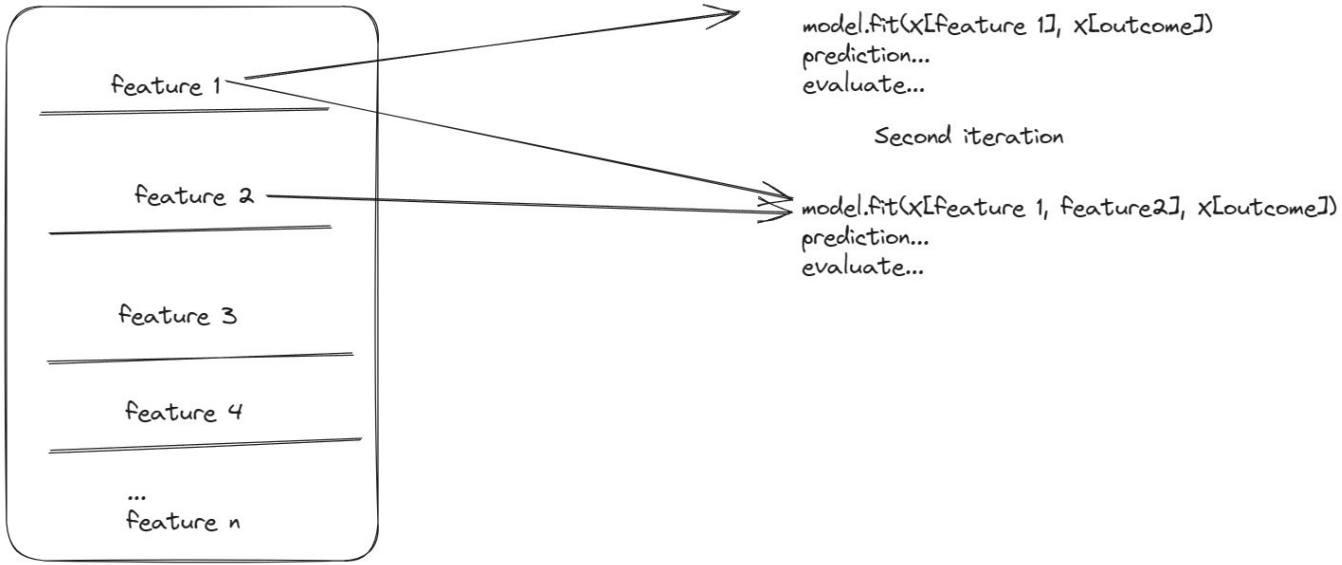
```
('rmse = 44753.37738568247', 'r2 = 0.673603470021773')
```

If the top correlated features do not give us the best results.

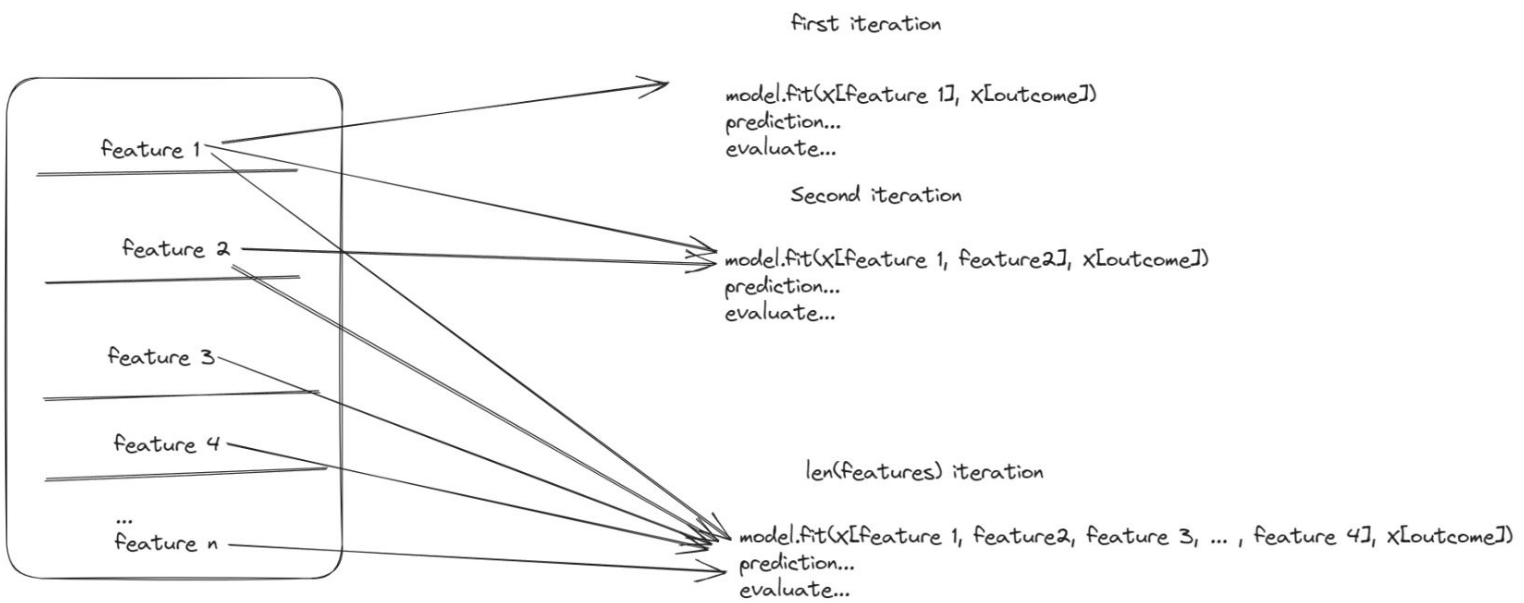
How do we find the features that would give us the best results?

I decided to brute force this process:

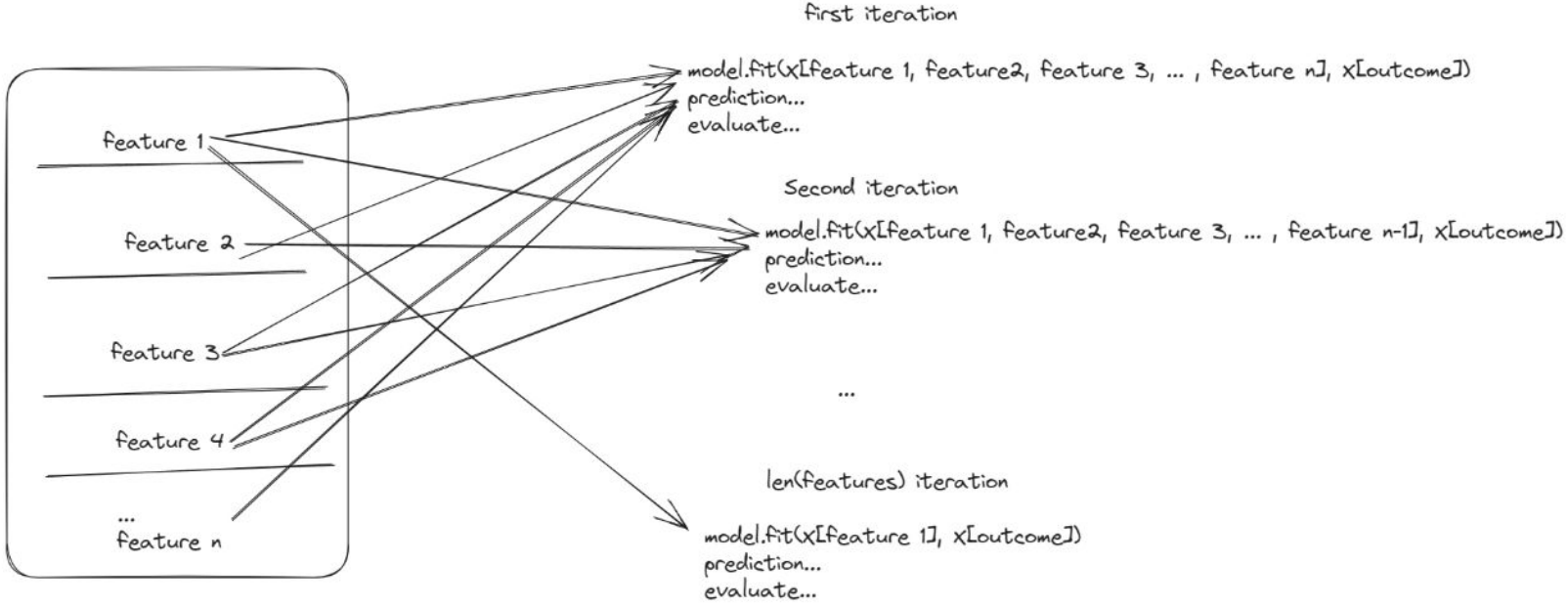
The forward selection algorithm



The forward selection algorithm



The Backward selection algorithm



Model Description	RMSE	R2	Initial # Features	Final # Features (if applicable)
Model of highest correlated features	40412.86100369427	0.7626541968542401	10	-
Model of highest correlated features with Cross validation	mean rmse = 36949.87866445592	mean r2 = 0.779848655366105	10	-
Model of highest correlated features with Lasso	44359.00408473574	0.7140395731173013	10	-
Model of all numeric features	38337.59165478559	0.7864045402331978	37	-
Model of all numeric features with Cross Validation	mean rmse = 35642.85184537606	mean r2 = 0.791738446278189	37	-
Model of all numeric features with Lasso	37934.24414094182	0.7908753474821981	37	-
Model of all features with dummied preprocessing	34243.128952623614	0.8295922881232749	209	-
Model of all features with dummied preprocessing and CV	mean rmse = 34290.61380036194	mean r2 = 0.8002326002217905	209	-
Model of all features with dummied preprocessing and Lasso	37934.24414094182	0.7908753474821981	209	-
Forward selection algorithm with numeric features only	37753.02406324579	0.7928686421507606	37	33
Forward selection algorithm with numeric and dummied features	33840.37573779651	0.8335772418404599	209	120
Backward selection algorithm with numeric features only	41751.280747375204	0.7466727159535904	209	36
Backward selection algorithm with numeric and dummied features	35439.69000658719	0.8174750696856952	209	208

Forward selection algorithm with numeric and dummied features	33840.37573779651	0.8335772418404599	209	120
Backward selection algorithm with numeric features only	41751.280747375204	0.7466727159535904	209	36
Backward selection algorithm with numeric and dummied features	35439.69000658719	0.8174750696856952	209	208

Summary of experiments with polynomial features

Model Description (features converted to polynomial of degree 2)	**RMSE**	**R2**	**Initial # Features**
Model with all numeric features and grid searching for best alpha	38505.77372194258	0.7845263982307846	37
Model with LassoCV with numeric features only	38421.485445693455	0.7854686995389903	37
Model with LassoCV with numeric and dummied	36777.60317703496	0.803433634559926	209
Model with all numeric features and *NaN filled with mean of data distribution*	23139.72146085574	0.9104218737234883	39

The best model ?

Simple Linear regression trained on the optimal features.

Optimal features were found using *stepwise algorithm*.

Stepwise algorithm uses forward and backward selection simultaneously.

Results:

Best performing and Final Model	RMSE	R2	Initial # Features	Final # Features (if applicable)
Model with all features trained using stepwise selection algorithm with simple linear regression	19455.7677	0.938694	221	94

Conclusion:

The results form the stepwise algorithm proves the fact that simplest linear regression is the best if all optimal features are weighed in.

Thank You !