Data Mining HW 3

Kaushal Nerkar

Task 1

Normalized Data Attributes

Train

```
1 df_train_norm_filtered
```

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	...	WoodDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch	ScreenPorch	PoolArea	MiscVal	MoSold	YrSold	SalePrice
0	1	120	30.0	9549	8	5	1995	1996	0.0	437	...	0	30	0	0	216	0	0	4	2006	270000
1	2	45	60.0	9000	6	3	1928	1950	0.0	0	...	0	0	91	0	0	0	0	10	2009	76000
2	3	60	NaN	12394	7	5	2003	2003	0.0	0	...	100	48	0	0	0	0	0	10	2007	225000
3	4	20	75.0	11957	8	5	2006	2006	53.0	24	...	144	104	0	0	0	0	0	7	2008	232000
4	5	20	NaN	11616	5	5	1962	1962	116.0	170	...	0	20	144	0	0	0	0	9	2009	139000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	996	50	50.0	6000	4	6	1936	1950	0.0	672	...	0	0	0	0	0	0	0	7	2009	128000
996	997	20	60.0	6960	4	6	1970	1970	0.0	375	...	96	0	0	0	0	0	500	11	2009	120500
997	998	20	78.0	10206	8	5	2007	2007	468.0	33	...	144	99	0	0	0	0	0	10	2008	245000
998	999	60	NaN	11214	7	5	1998	1999	0.0	0	...	89	0	0	0	0	0	0	7	2006	199900
999	1000	190	75.0	11625	5	4	1965	1965	0.0	841	...	0	0	0	0	0	0	0	4	2010	131500

1000 rows × 38 columns

Test

```
1 df_test_norm_filtered
```

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	...	WoodDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch	ScreenPorch	PoolArea	MiscVal	MoSold	YrSold	SalePrice
0	1	20	80.0	11900	6	5	1957	1957	387.0	1040	...	0	0	0	0	192	0	0	6	2008	166000
1	2	90	78.0	7060	7	5	1997	1998	200.0	1309	...	0	0	0	0	0	0	0	11	2008	206300
2	3	60	NaN	7851	6	5	2002	2002	NaN	625	...	288	48	0	0	0	0	0	5	2010	216500
3	4	60	NaN	11000	8	5	2000	2000	72.0	0	...	0	56	0	0	0	0	0	6	2007	248000
4	5	180	35.0	3675	6	5	2005	2005	80.0	459	...	0	28	0	0	0	0	0	6	2008	148000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
455	456	190	70.0	7000	5	7	1910	1991	0.0	969	...	85	0	148	0	0	0	0	1	2008	107500
456	457	50	75.0	45600	6	8	1908	1997	0.0	0	...	486	40	0	0	175	0	0	9	2008	240000
457	458	20	NaN	11717	6	6	1970	1970	571.0	0	...	371	0	0	0	0	0	0	2	2009	185000
458	459	70	66.0	9042	7	9	1941	2006	0.0	275	...	0	60	0	0	0	0	2500	5	2010	266500
459	460	80	80.0	9600	6	6	1955	1996	0.0	831	...	0	28	0	0	178	0	0	10	2008	165500

460 rows × 38 columns

Selected Features: ['OverallQual', 'GrLivArea', 'BsmtFinSF1', 'GarageArea', 'YearRemodAdd', 'MSSubClass', 'MasVnrArea', 'Fireplaces', 'TotalBsmtSF', 'YearBuilt']

Sample Train data with selected features

```
1 X_train
```

	OverallQual	GrLivArea	BsmtFinSF1	GarageArea	YearRemodAdd	MSSubClass	MasVnrArea	Fireplaces	TotalBsmtSF	YearBuilt
0	0.045134	0.033232	0.025377	0.032285	0.036310	0.061385	0.000000	0.04222	0.046697	0.036535
1	0.033850	0.017439	0.000000	0.024163	0.035473	0.023019	0.000000	0.00000	0.024505	0.035308
2	0.045134	0.035011	0.001394	0.055307	0.036492	0.010231	0.008637	0.04222	0.049197	0.036737
3	0.039492	0.033921	0.081298	0.036312	0.036383	0.010231	0.039763	0.00000	0.047666	0.036627
4	0.039492	0.036257	0.000000	0.030271	0.036401	0.010231	0.027866	0.04222	0.050948	0.036627
...	...	...	...	...	...	...	...	...	...	...
761	0.016925	0.023133	0.000000	0.026848	0.035473	0.046039	0.000000	0.00000	0.000000	0.035711
762	0.022567	0.029450	0.039023	0.016109	0.035473	0.025577	0.000000	0.00000	0.021004	0.035455
763	0.022567	0.019218	0.021776	0.044300	0.035837	0.010231	0.000000	0.00000	0.027005	0.036078
764	0.045134	0.034766	0.001916	0.050877	0.036510	0.010231	0.076266	0.04222	0.048853	0.036755
765	0.028209	0.023111	0.048837	0.033829	0.035746	0.097193	0.000000	0.00000	0.032475	0.035986

766 rows × 10 columns

```
Mean Squared Error (MSE) Train Error:  3.826873504952882e-05
Mean Squared Error (MSE) Test Error: 0.009589233903404918
```

Task 2

5th degree Polynomial Transform

```
X_train shape before -> after  transform (766, 10) --> (766, 3003)
X_test  shape before -> after transform (355, 10) --> (355, 3003)
```

Task 3

```
Train Error(MSE) on Polynomially transformed data(deg=5) : 1.0209895058632275e-08
Test Error (MSE) on Polynomially Transformed data(deg=5) : 184.51528357862554
```

As we can see train error is low, but test error is increasing so we can say overfitting has occurred and we need to regularize it to fit model well.

Task 4

values for the regularization parameter

```
reg_values = [0.001, 0.01, 0.1, 1, 5, 10, 20, 50, 100, 1000]
```

```
///-----------------------------------Reg_parameter and ECv ----------------------------------------///
          Regularization Parameter : 0.001,  Cross-Validation Error: 0.000043063
          Regularization Parameter : 0.01,  Cross-Validation Error: 0.000044918
          Regularization Parameter : 0.1,  Cross-Validation Error: 0.000061588
          Regularization Parameter : 1,  Cross-Validation Error: 0.000123446
          Regularization Parameter : 5,  Cross-Validation Error: 0.000186151
          Regularization Parameter : 10,  Cross-Validation Error: 0.000202848
          Regularization Parameter : 20,  Cross-Validation Error: 0.000212799
          Regularization Parameter : 50,  Cross-Validation Error: 0.000219395
          Regularization Parameter : 100,  Cross-Validation Error: 0.000221710
          Regularization Parameter : 1000,  Cross-Validation Error: 0.000223845
    ///-----------------------------------------------------------------------------------------------------
    Selected Value of Regularization Paramneter : 10
    Training Error with Reg_p 10 is : 0.00019980752299344337
    Test Error with Reg_p 10 is : 0.0006204489701553935
```

Task 5

Chosen MLP parameters.

```
parameters = {
    'hidden_layer_sizes': [(50,), (100,), (50, 50), (100, 100)],
    'alpha': [0.0001, 0.001, 0.01],
}
```

```
Mean cross-validation Error for all models:
ECv -0.000371012, Parameters: {'alpha': 0.0001, 'hidden_layer_sizes': (50,)}
ECv -0.000195479, Parameters: {'alpha': 0.0001, 'hidden_layer_sizes': (100,)}
ECv -0.000202535, Parameters: {'alpha': 0.0001, 'hidden_layer_sizes': (50, 50)}
ECv -0.000141606, Parameters: {'alpha': 0.0001, 'hidden_layer_sizes': (100, 100)}
ECv -0.000271681, Parameters: {'alpha': 0.001, 'hidden_layer_sizes': (50,)}
ECv -0.000196039, Parameters: {'alpha': 0.001, 'hidden_layer_sizes': (100,)}
ECv -0.000158889, Parameters: {'alpha': 0.001, 'hidden_layer_sizes': (50, 50)}
ECv -0.000113189, Parameters: {'alpha': 0.001, 'hidden_layer_sizes': (100, 100)}
ECv -0.000311873, Parameters: {'alpha': 0.01, 'hidden_layer_sizes': (50,)}
ECv -0.000199820, Parameters: {'alpha': 0.01, 'hidden_layer_sizes': (100,)}
ECv -0.000215858, Parameters: {'alpha': 0.01, 'hidden_layer_sizes': (50, 50)}
ECv -0.000105994, Parameters: {'alpha': 0.01, 'hidden_layer_sizes': (100, 100)}
```

```
Optimal Model Parameters  :  {'alpha': 0.01, 'hidden_layer_sizes': (100, 100)}
Best cross-validation Error (negative MSE): -0.00010599390884324517
Train MSE with regp 0.01 is : 7.732782172785152e-05
Test MSE with regp 0.01 is: 0.00026736394202103806
```

Task 6

Train data with added Attribute Class (1, -1) based on Sale Price

```
1 df_train_class
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice	Class
0	1	120	RL	30.0	9549	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	NaN	0	4	2006	WD	Normal	270000	-1
1	2	45	RH	60.0	9000	Pave	NaN	Reg	Lvl	AllPub	...	NaN	NaN	NaN	0	10	2009	WD	Normal	76000	1
2	3	60	RL	NaN	12394	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	NaN	0	10	2007	WD	Family	225000	-1
3	4	20	RL	75.0	11957	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	NaN	0	7	2008	WD	Normal	232000	-1
4	5	20	RL	NaN	11616	Pave	NaN	Reg	Lvl	AllPub	...	NaN	NaN	NaN	0	9	2009	WD	Normal	139000	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	996	50	RM	50.0	6000	Pave	NaN	Reg	Lvl	AllPub	...	NaN	NaN	NaN	0	7	2009	WD	Normal	128000	1
996	997	20	RL	60.0	6960	Pave	NaN	Reg	Lvl	AllPub	...	NaN	NaN	Shed	500	11	2009	WD	Normal	120500	1
997	998	20	RL	78.0	10206	Pave	NaN	Reg	Lvl	AllPub	...	NaN	NaN	NaN	0	10	2008	WD	Normal	245000	-1
998	999	60	RL	NaN	11214	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	NaN	0	7	2006	WD	Normal	199900	-1
999	1000	190	RL	75.0	11625	Pave	NaN	Reg	Lvl	AllPub	...	NaN	NaN	NaN	0	4	2010	WD	Normal	131500	1

1000 rows × 82 columns

Test data with added Attribute Class (1, -1) based on Sale Price

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	Class
0	1	20	RL	80.0	11900	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	6	2008	WD	Normal	-1
1	2	90	RM	78.0	7060	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	11	2008	WD	Alloca	-1
2	3	60	RL	NaN	7851	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	5	2010	WD	Normal	-1
3	4	60	RL	NaN	11000	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	6	2007	WD	Normal	-1
4	5	180	RM	35.0	3675	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	6	2008	WD	Normal	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
455	456	190	RL	70.0	7000	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	1	2008	WD	Normal	1
456	457	50	RL	75.0	45600	Pave	NaN	IR2	Bnk	AllPub	...	0	NaN	NaN	NaN	0	9	2008	WD	Normal	-1
457	458	20	RL	NaN	11717	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2009	WD	Normal	-1
458	459	70	RL	66.0	9042	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	GdPrv	Shed	2500	5	2010	WD	Normal	-1
459	460	80	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0	10	2008	WD	Normal	-1

460 rows × 81 columns

Link to Colab:
https://colab.research.google.com/drive/1IhgZP0uGcRMmNj6GKE7wLkO3tgDom4Tt?authuser=0#scrollTo=qojY7chzPXtH&uniqifier=2