

# Chapter 1: Get Closer to Your Data

	Id	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	BsmtFinSF2	BsmtUnfSF
count	1460.000000	1201.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1452.000000	1460.000000	1460.000000	1460.000000
mean	730.500000	70.049958	10516.828082	6.099315	5.575342	1971.267808	1984.865753	103.685262	443.639726	46.549315	567.240411
std	421.610009	24.284752	9981.264932	1.382997	1.112799	30.202904	20.645407	181.066207	456.098091	161.319273	441.866955
min	1.000000	21.000000	1300.000000	1.000000	1.000000	1872.000000	1950.000000	0.000000	0.000000	0.000000	0.000000
25%	365.750000	59.000000	7553.500000	5.000000	5.000000	1954.000000	1967.000000	0.000000	0.000000	0.000000	223.000000
50%	730.500000	69.000000	9478.500000	6.000000	5.000000	1973.000000	1994.000000	0.000000	383.500000	0.000000	477.500000
75%	1095.250000	80.000000	11601.500000	7.000000	6.000000	2000.000000	2004.000000	166.000000	712.250000	0.000000	808.000000
max	1460.000000	313.000000	215245.000000	10.000000	9.000000	2010.000000	2010.000000	1600.000000	5644.000000	1474.000000	2336.000000

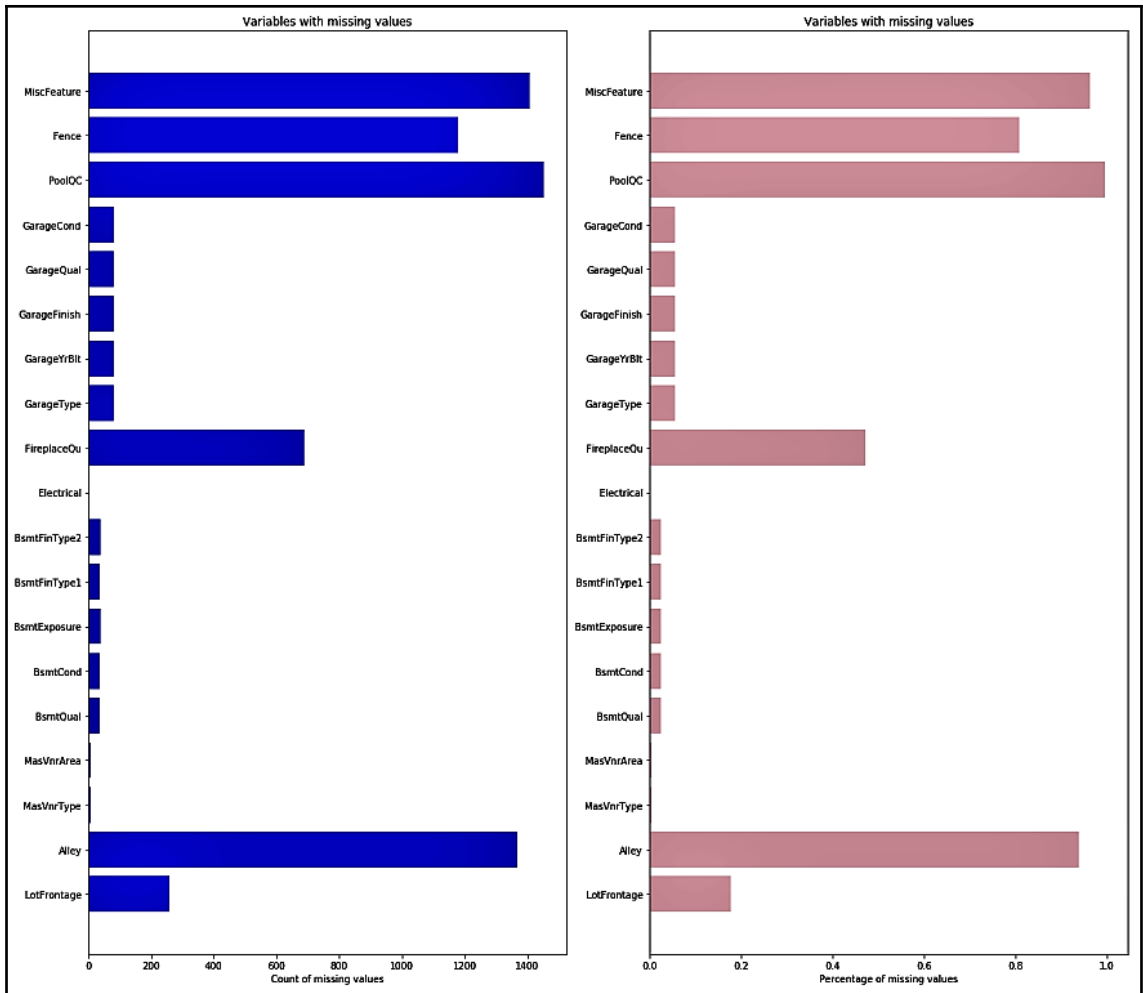
Utilities	...	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice	building_age	remodelled_age
AllPub	...	NaN	NaN	0	2	2008	WD	Normal	208500	15	15
AllPub	...	NaN	NaN	0	5	2007	WD	Normal	181500	42	42
AllPub	...	NaN	NaN	0	9	2008	WD	Normal	223500	17	16
AllPub	...	NaN	NaN	0	2	2006	WD	Abnorml	140000	103	48
AllPub	...	NaN	NaN	0	12	2008	WD	Normal	250000	18	18

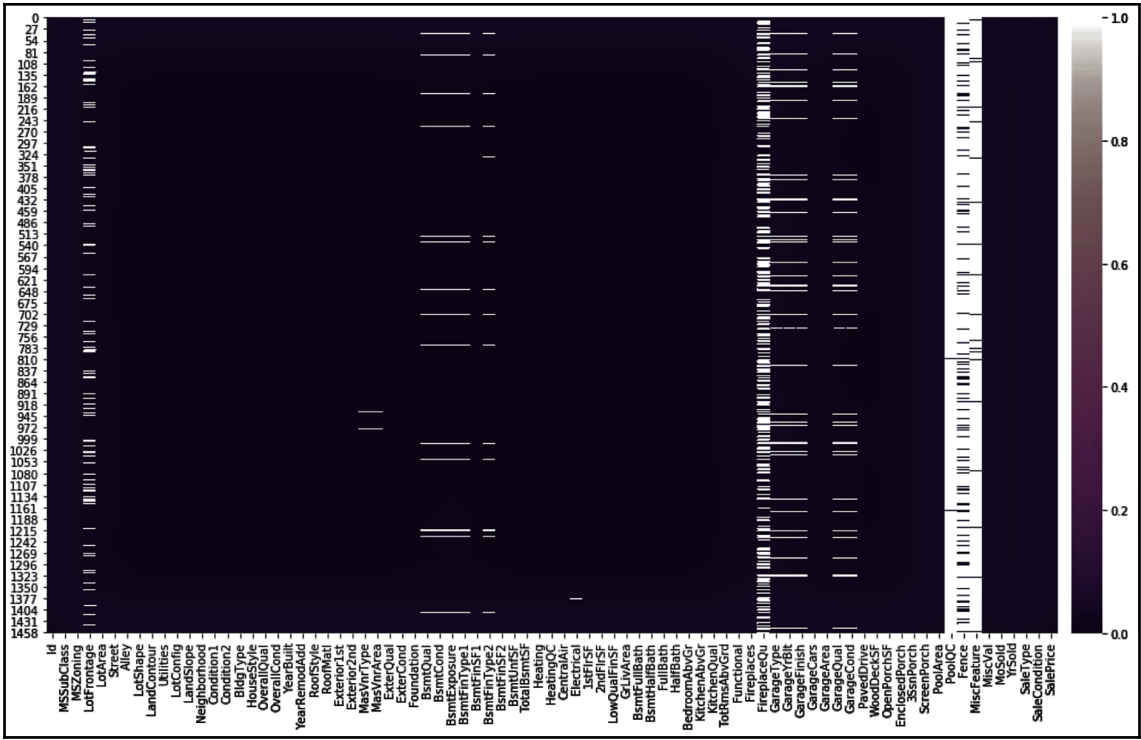
	LotShape_IR1	LotShape_IR2	LotShape_IR3	LotShape_Reg
0	0	0	0	1
1	0	0	0	1
2	1	0	0	0
3	1	0	0	0
4	1	0	0	0
5	1	0	0	0

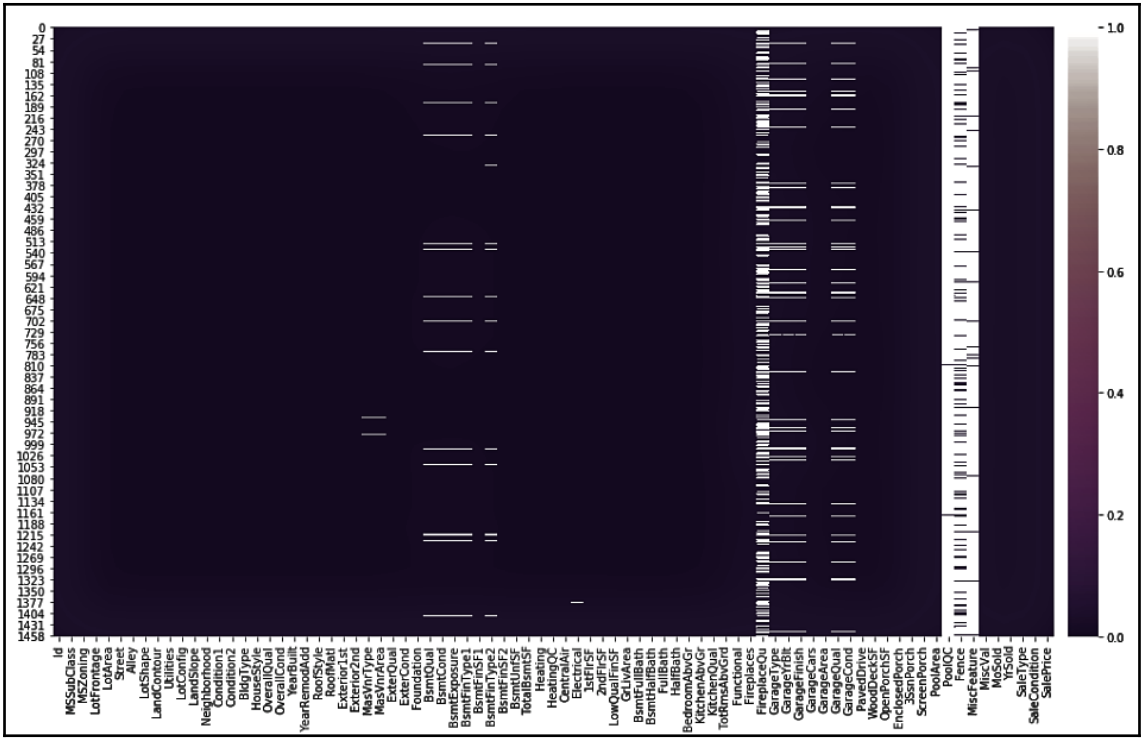
SaleType	SaleCondition	SalePrice	building_age	remodelled_age	LotShape_IR1	LotShape_IR2	LotShape_IR3	LotShape_Reg
WD	Normal	208500	15	15	0	0	0	1
WD	Normal	181500	42	42	0	0	0	1
WD	Normal	223500	17	16	1	0	0	0
WD	Abnorml	140000	103	48	1	0	0	0
WD	Normal	250000	18	18	1	0	0	0

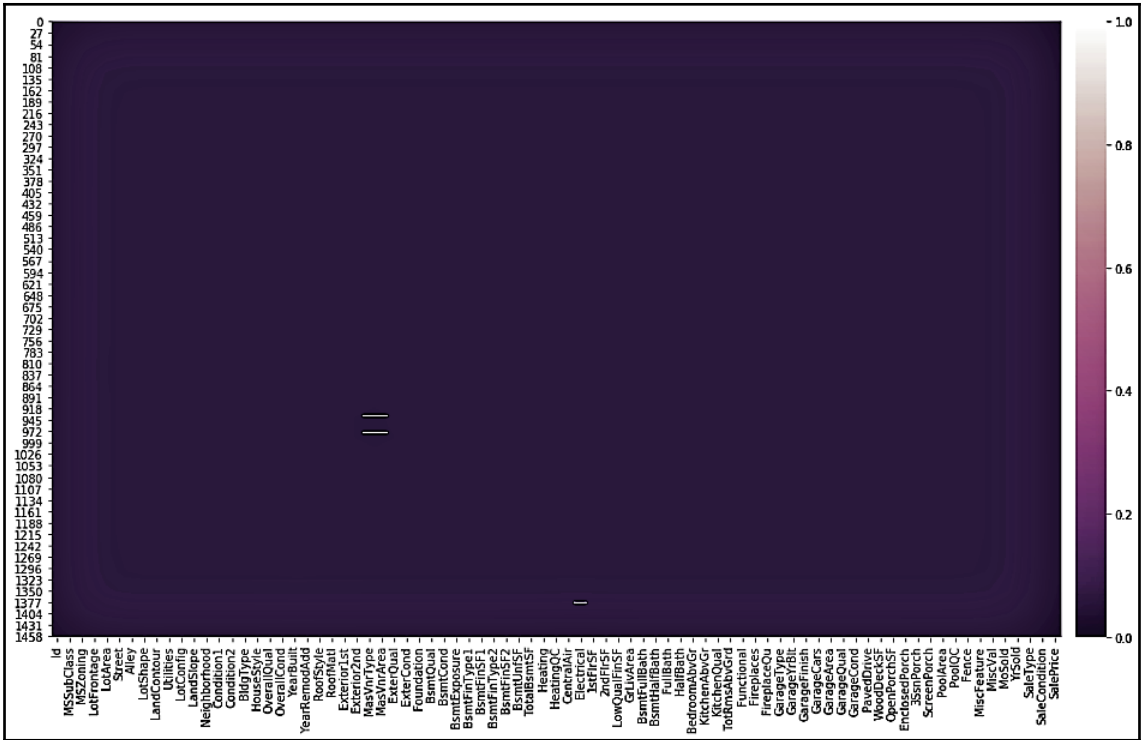
---

LotFrontage	259
Alley	1369
MasVnrType	8
MasVnrArea	8
BsmtQual	37
BsmtCond	37
BsmtExposure	38
BsmtFinType1	37
BsmtFinType2	38
Electrical	1
FireplaceQu	690
GarageType	81
GarageYrBlt	81
GarageFinish	81
GarageQual	81
GarageCond	81
PoolQC	1453
Fence	1179
MiscFeature	1406









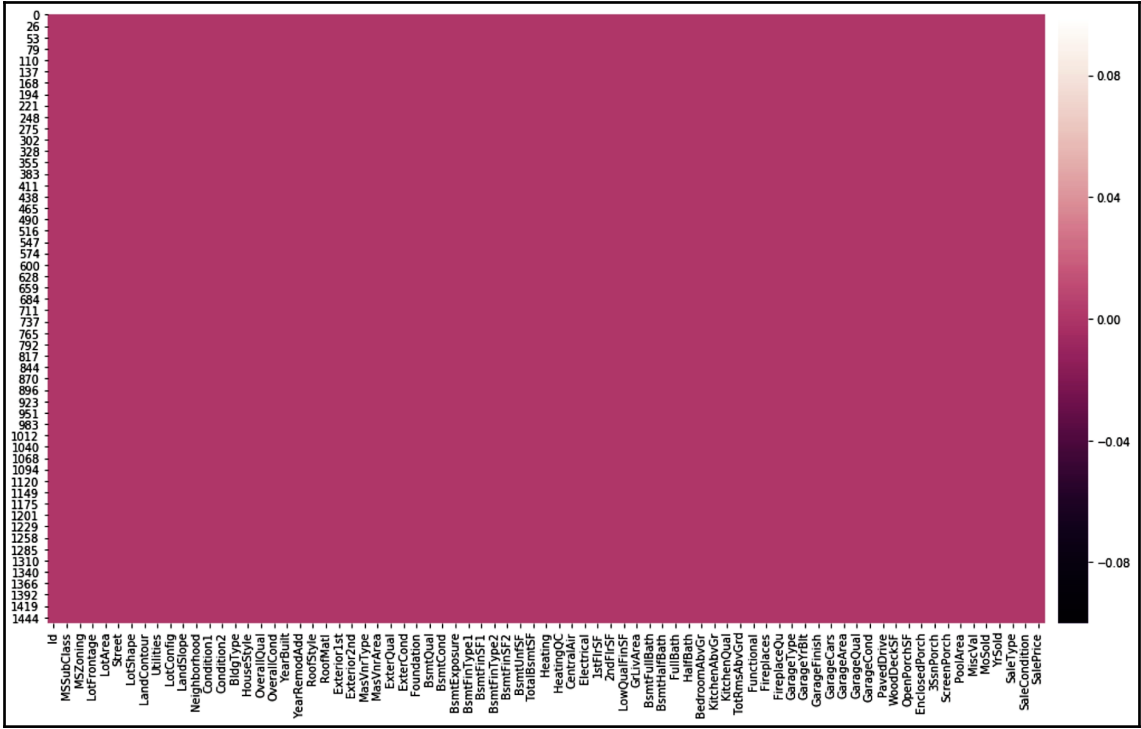
MasVnrArea	0.0	1.0	11.0	14.0	16.0	18.0	22.0	24.0	27.0	28.0	30.0
MasVnrType											
BrkCmn	0	0	0	0	0	0	0	0	0	0	0
BrkFace	1	0	1	1	6	2	1	0	1	1	2
None	859	2	0	0	0	0	0	0	0	0	0
Stone	1	0	0	0	1	0	0	1	0	0	0
All	861	2	1	1	7	2	1	1	1	1	2

```

1379      80
Name: MSSubClass, dtype: int64

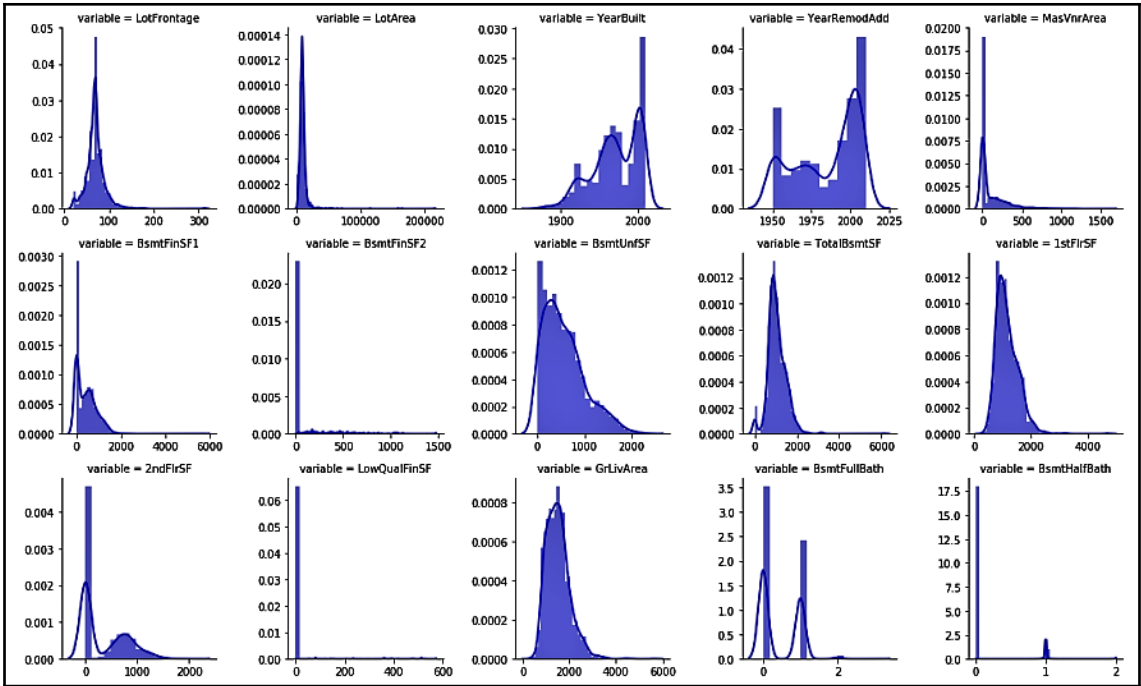
```

MSSubClass	20	30	40	45	50	60	70	75	80	85	90	120	160	180	190
Electrical															
FuseA	31	18	2	3	20	0	8	2	0	0	5	0	0	0	5
FuseF	5	4	0	2	8	0	2	0	1	0	4	0	0	0	1
FuseP	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1
Mix	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
SBrkr	500	45	2	7	116	299	50	14	56	20	42	87	63	10	23
All	536	69	4	12	144	299	60	16	58	20	52	87	63	10	30

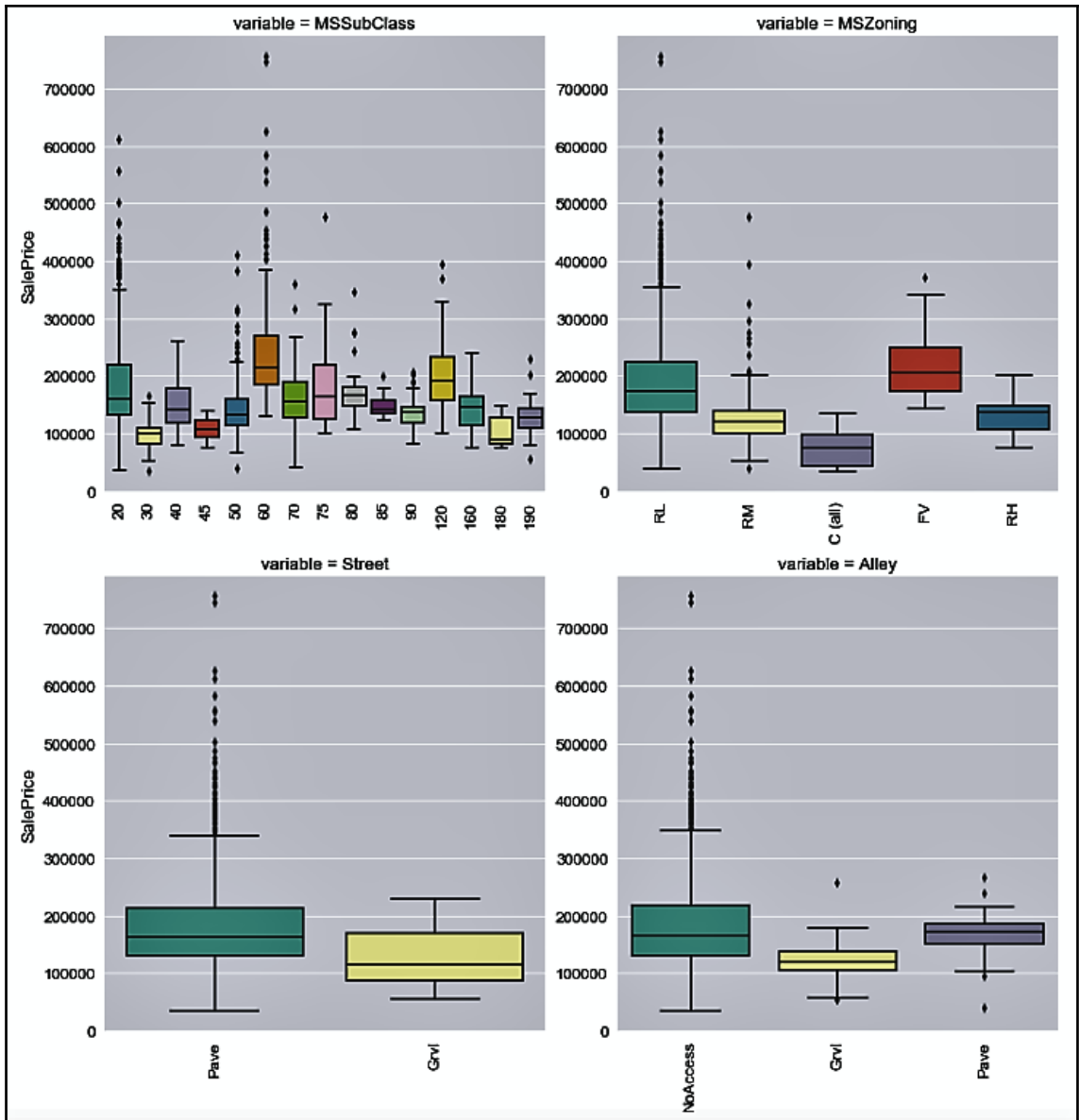


```
float64      3
int64       33
object       45
uint8        4
dtype: int64
```

```
Number of Numerical features: 40
Number of Categorical features: 45
```

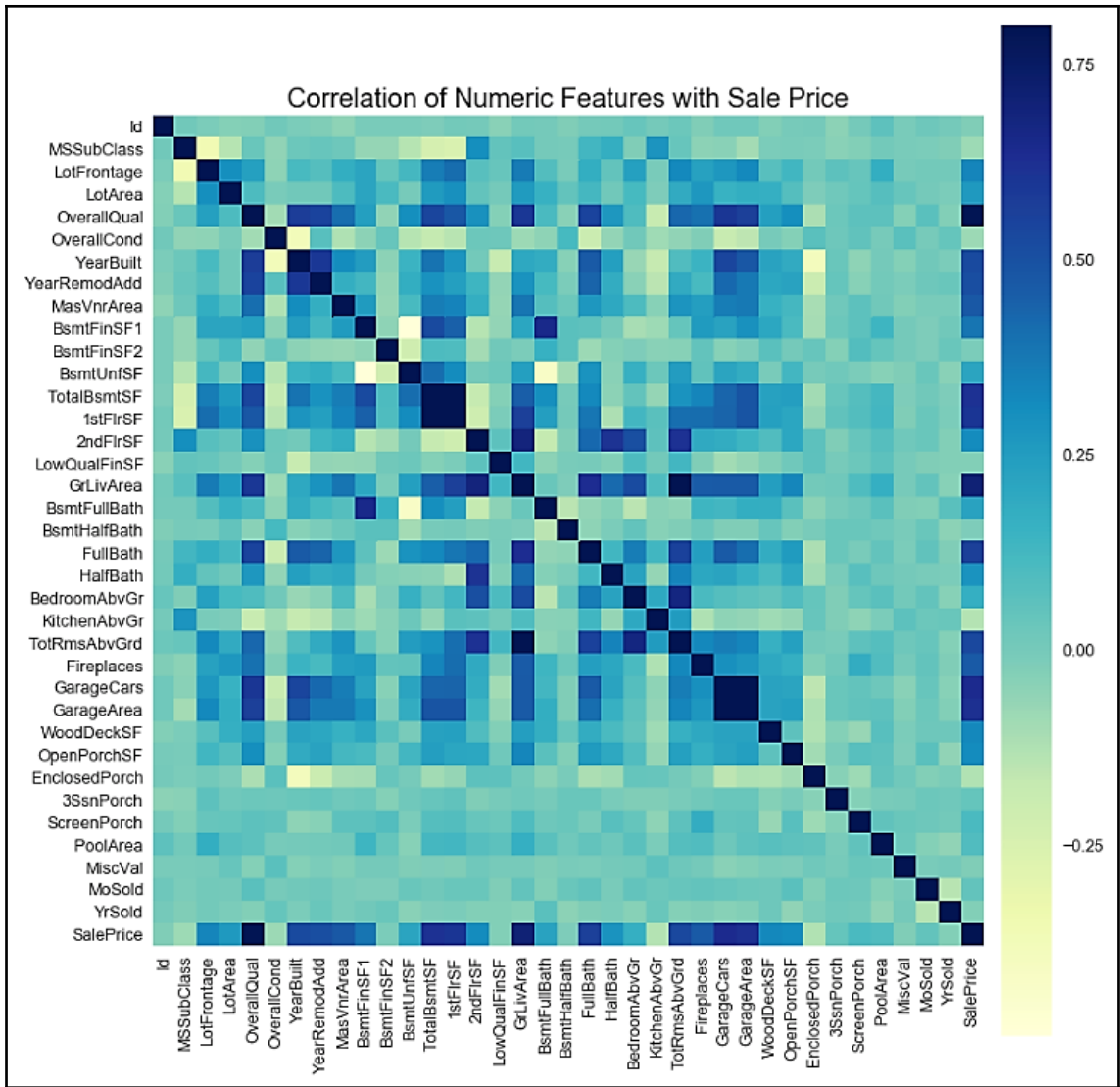


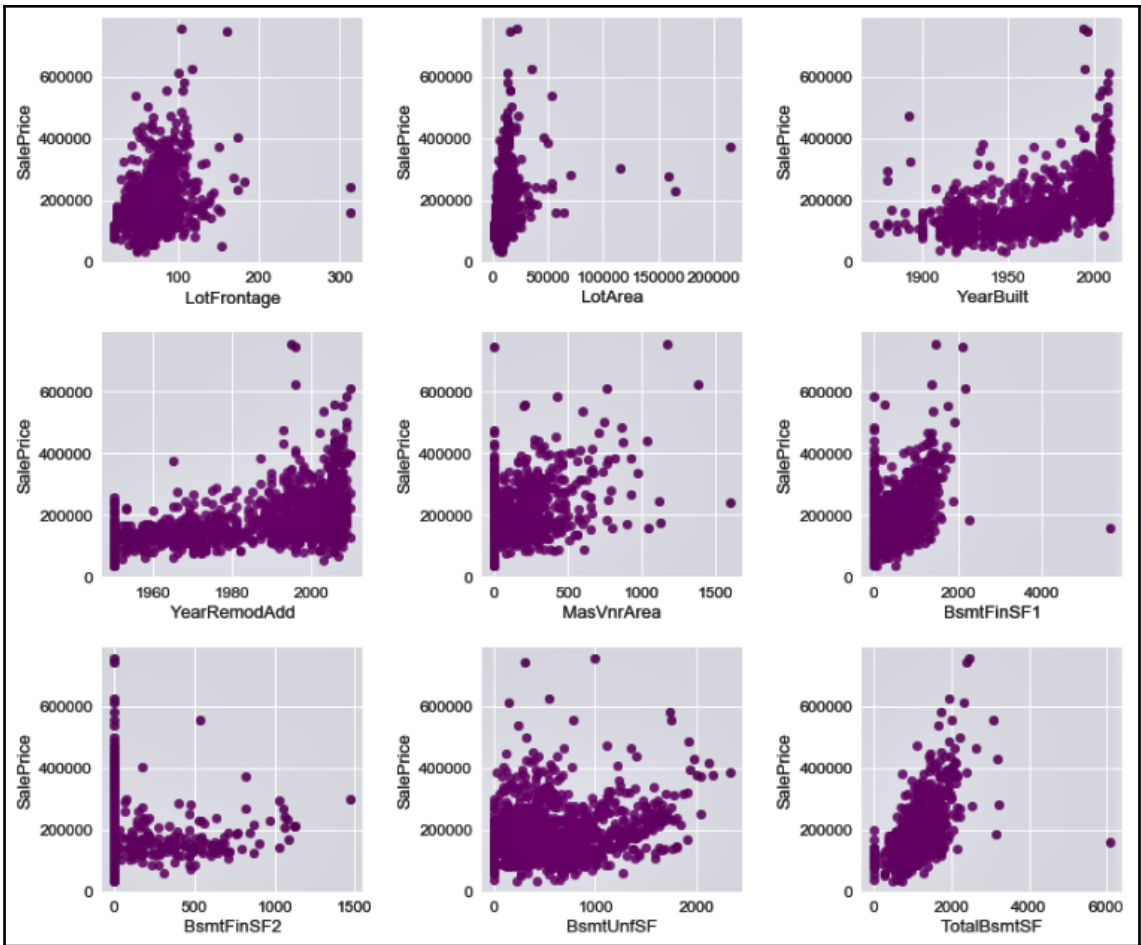




---

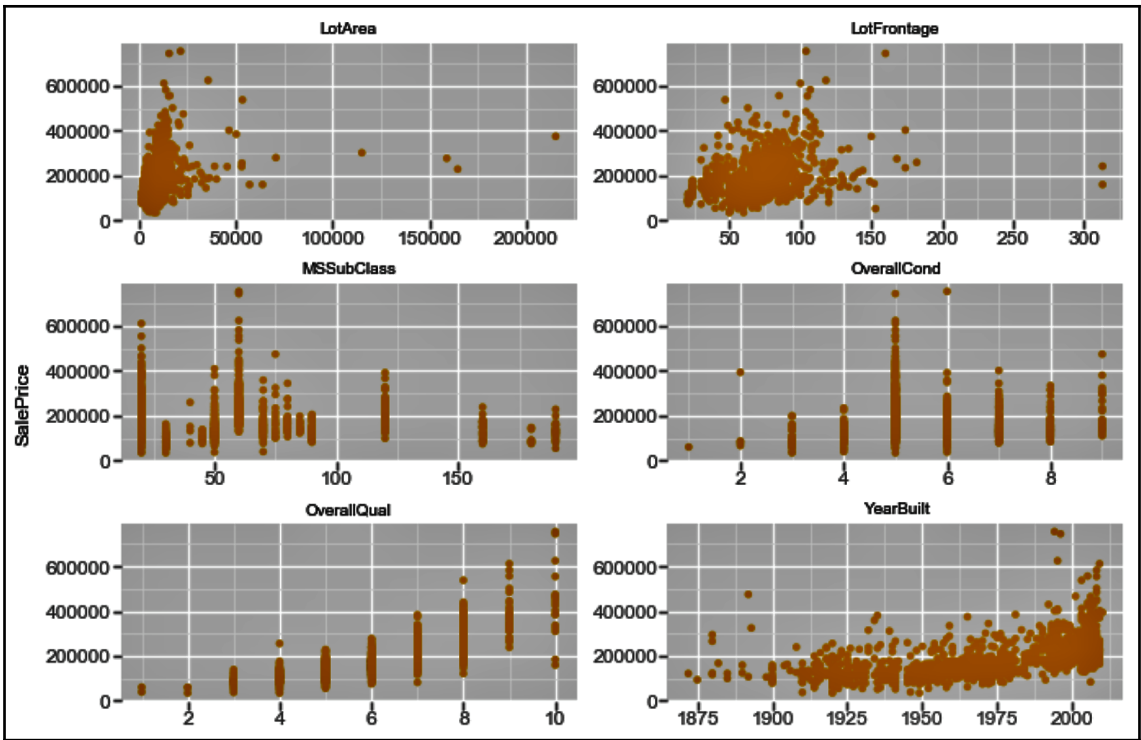
	LotFrontage	LotArea	YearBuilt	YearRemodAdd	MasVnrArea
LotFrontage	1.000000	0.304522	0.116685	0.083348	0.178469
LotArea	0.304522	1.000000	0.014228	0.013788	0.103321
YearBuilt	0.116685	0.014228	1.000000	0.592855	0.311600
YearRemodAdd	0.083348	0.013788	0.592855	1.000000	0.176529
MasVnrArea	0.178469	0.103321	0.311600	0.176529	1.000000
BsmtFinSF1	0.214367	0.214103	0.249503	0.128451	0.261256
BsmtFinSF2	0.042463	0.111170	-0.049107	-0.067759	-0.071330
BsmtUnfSF	0.124098	-0.002618	0.149040	0.181133	0.113862
TotalBsmtSF	0.363472	0.260833	0.391452	0.291066	0.360067
1stFlrSF	0.413773	0.299475	0.281986	0.240379	0.339850
2ndFlrSF	0.072388	0.050986	0.010308	0.140024	0.173800
LowQualFinSF	0.037469	0.004779	-0.183784	-0.062419	-0.068628
GrLivArea	0.368007	0.263116	0.199010	0.287389	0.388052

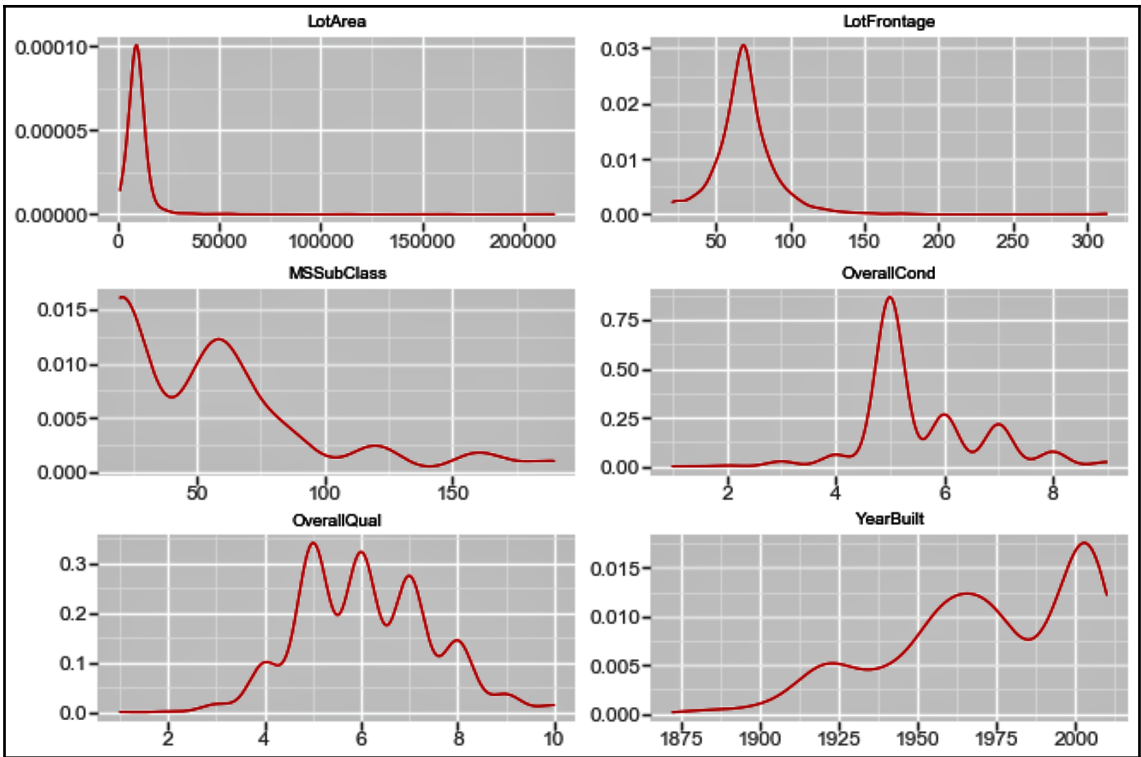




---

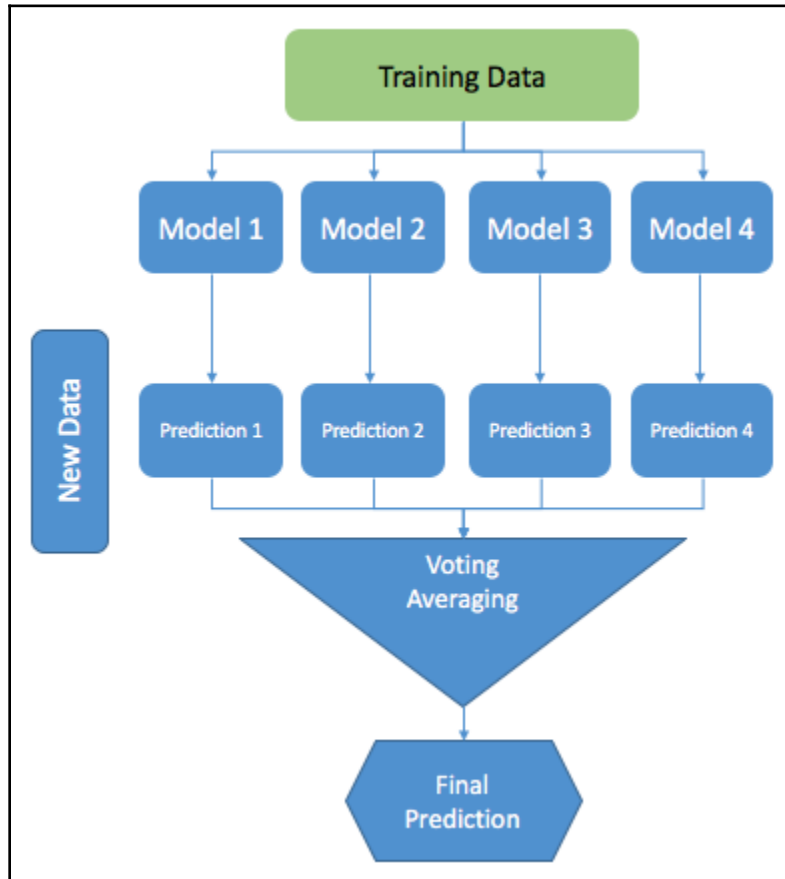
SalePrice	1.000000
GrLivArea	0.708624
GarageCars	0.640409
GarageArea	0.623431
TotalBsmtSF	0.613581
1stFlrSF	0.605852
FullBath	0.560664
TotRmsAbvGrd	0.533723
YearBuilt	0.522897
YearRemodAdd	0.507101
MasVnrArea	0.472614
Fireplaces	0.466929
BsmtFinSF1	0.386420
LotFrontage	0.334771
WoodDeckSF	0.324413
2ndFlrSF	0.319334
OpenPorchSF	0.315856
HalfBath	0.284108
LotArea	0.263843
BsmtFullBath	0.227122
BsmtUnfSF	0.214479
BedroomAbvGr	0.168213
ScreenPorch	0.111447
PoolArea	0.092404
MoSold	0.046432
3SsnPorch	0.044584
BsmtFinSF2	-0.011378
BsmtHalfBath	-0.016844
MiscVal	-0.021190
LowQualFinSF	-0.025606
YrSold	-0.028923
EnclosedPorch	-0.128578
KitchenAbvGr	-0.135907





---

## Chapter 2: Getting Started with Ensemble Machine Learning



$$MSE = \underbrace{[E[\hat{f}(x)] - f(x)]^2}_{\text{Bias}} + \underbrace{E[(\hat{f}(x) - E[\hat{f}(x)])^2]}_{\text{Variance}} + \underbrace{\epsilon}_{\text{Noise}}$$



---

```
DecisionTreeClassifier 0.833333333333
SVC 0.944444444444
LogisticRegression 0.777777777778
```

```
Classifier Accuracy using Hard Voting: 0.944444444444
```

```
DecisionTreeClassifier 0.833333333333
SVC 0.944444444444
LogisticRegression 0.777777777778
Classifier Accuracy using Soft Voting: 0.888888888889
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	6.7	0.62	0.24	1.10	0.039	6.0	62.0	0.99340	3.41	0.32	10.400000	5
1	5.7	0.22	0.20	16.00	0.044	41.0	113.0	0.99862	3.22	0.46	8.900000	6
2	5.9	0.19	0.26	7.40	0.034	33.0	123.0	0.99500	3.49	0.42	10.100000	6
3	5.3	0.47	0.10	1.30	0.036	11.0	74.0	0.99082	3.48	0.54	11.200000	4
4	6.4	0.29	0.21	9.65	0.041	36.0	119.0	0.99334	2.99	0.34	10.933333	6

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
0	87139402	B	12.32	12.39	78.85	464.1	0.10280	0.06981
1	8910251	B	10.60	18.95	69.28	346.4	0.09688	0.11470
2	905520	B	11.04	16.83	70.92	373.2	0.10770	0.07804
3	868871	B	11.28	13.39	73.00	384.8	0.11640	0.11360
4	9012568	B	15.19	13.21	97.65	711.8	0.07963	0.06934

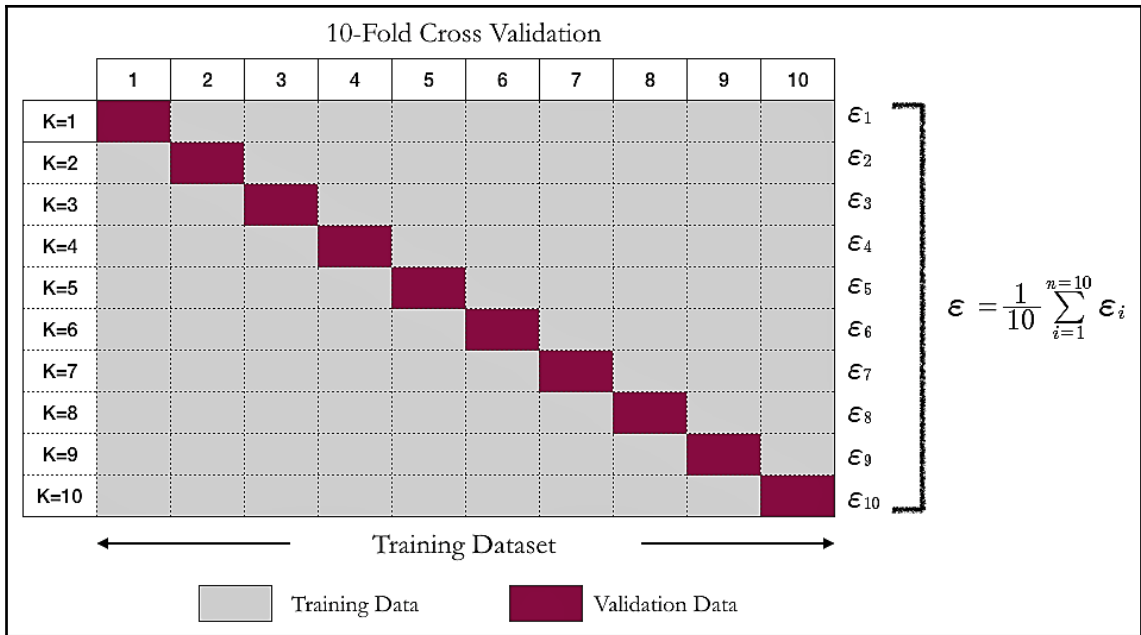
---

## Chapter 3: Resampling Methods

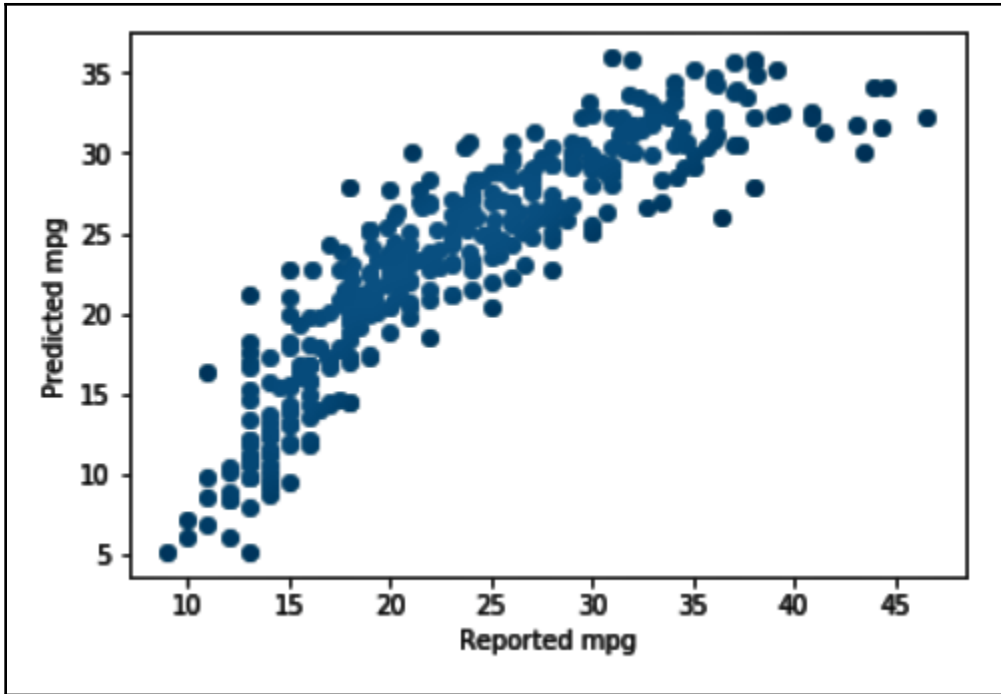
```
(1460, 80)
```

```
(1021, 80)  
(1021, )  
(438, 80)  
(438, )
```

```
0    77.88  
1    22.12  
dtype: float64  
0    77.88  
1    22.12  
dtype: float64
```

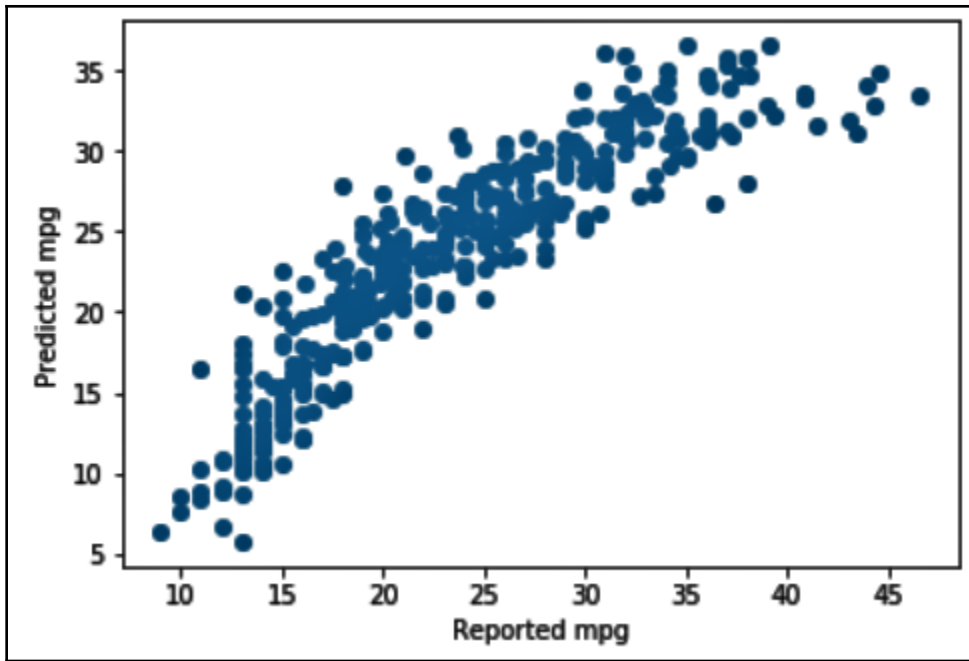


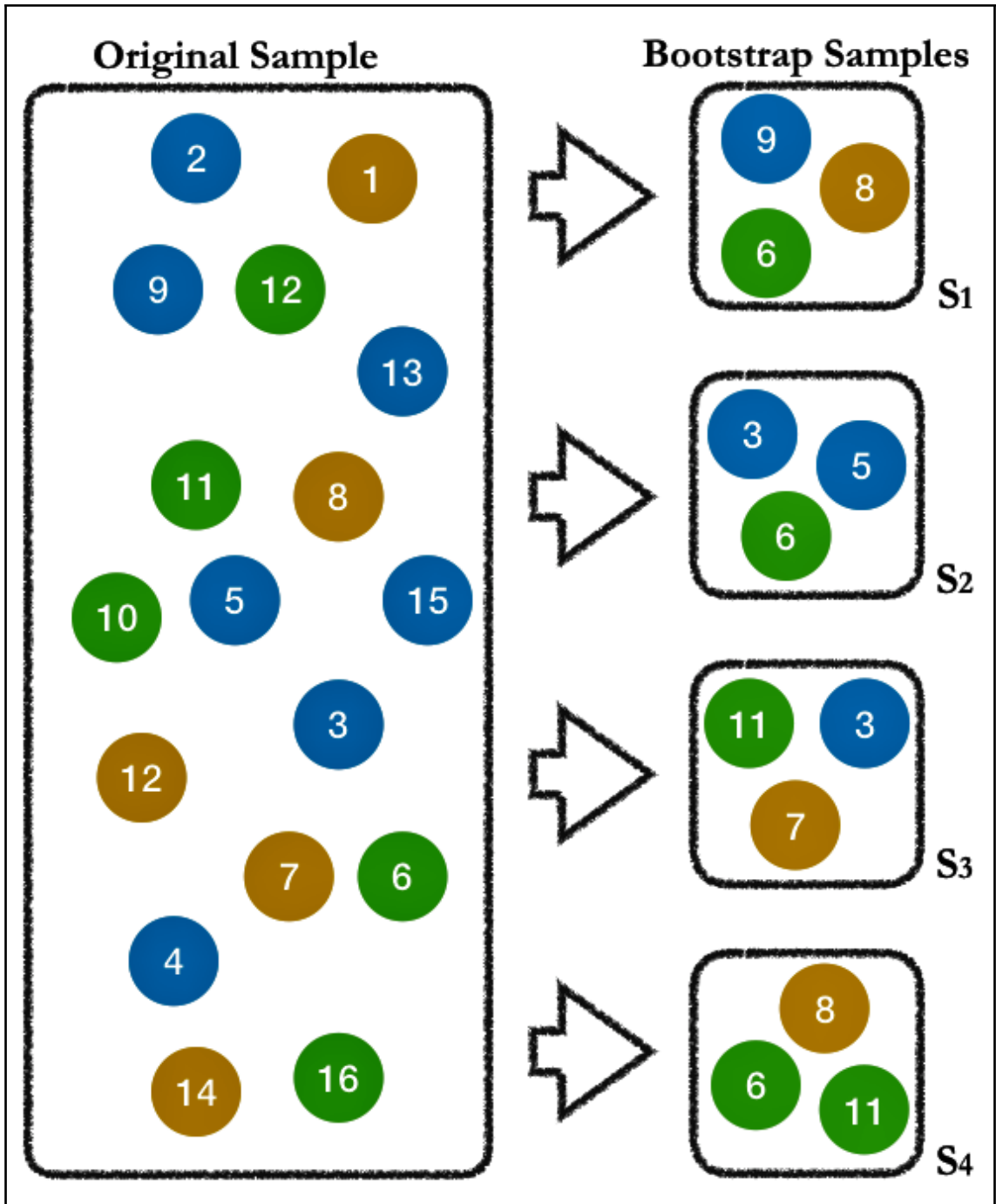
**R<sup>2</sup>: 0.68, MSE: 8.96**  
**R<sup>2</sup>: 0.73, MSE: 9.10**  
**R<sup>2</sup>: 0.69, MSE: 9.90**  
**R<sup>2</sup>: 0.72, MSE: 9.47**  
**R<sup>2</sup>: 0.75, MSE: 8.50**  
**R<sup>2</sup>: 0.77, MSE: 8.32**  
**R<sup>2</sup>: 0.75, MSE: 9.83**  
**R<sup>2</sup>: 0.77, MSE: 10.13**  
**R<sup>2</sup>: 0.78, MSE: 12.59**  
**R<sup>2</sup>: 0.79, MSE: 12.85**  
**Average CV Score : 9.962852981837424**

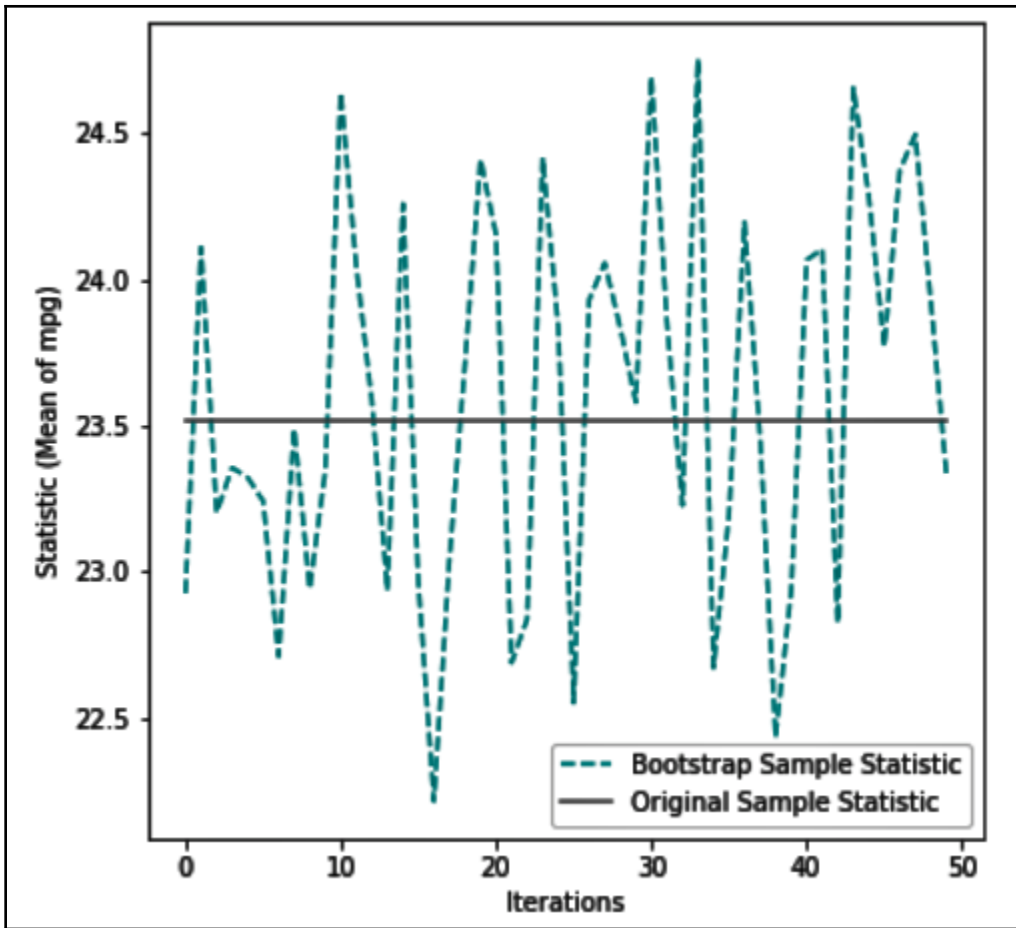


---

R <sup>2</sup> : 0.81, MSE: 11.11
R <sup>2</sup> : 0.81, MSE: 11.09
R <sup>2</sup> : 0.81, MSE: 11.06
R <sup>2</sup> : 0.81, MSE: 11.04
R <sup>2</sup> : 0.82, MSE: 11.02
R <sup>2</sup> : 0.82, MSE: 11.03
R <sup>2</sup> : 0.82, MSE: 11.02
R <sup>2</sup> : 0.82, MSE: 11.00
R <sup>2</sup> : 0.81, MSE: 11.23
R <sup>2</sup> : 0.81, MSE: 11.24
R <sup>2</sup> : 0.81, MSE: 11.32
R <sup>2</sup> : 0.81, MSE: 11.29
R <sup>2</sup> : 0.81, MSE: 11.33
R <sup>2</sup> : 0.81, MSE: 11.31
R <sup>2</sup> : 0.81, MSE: 11.28
R <sup>2</sup> : 0.81, MSE: 11.50
R <sup>2</sup> : 0.81, MSE: 11.47
R <sup>2</sup> : 0.81, MSE: 11.45
R <sup>2</sup> : 0.81, MSE: 11.43
Average CV Score : 8.401498720470839

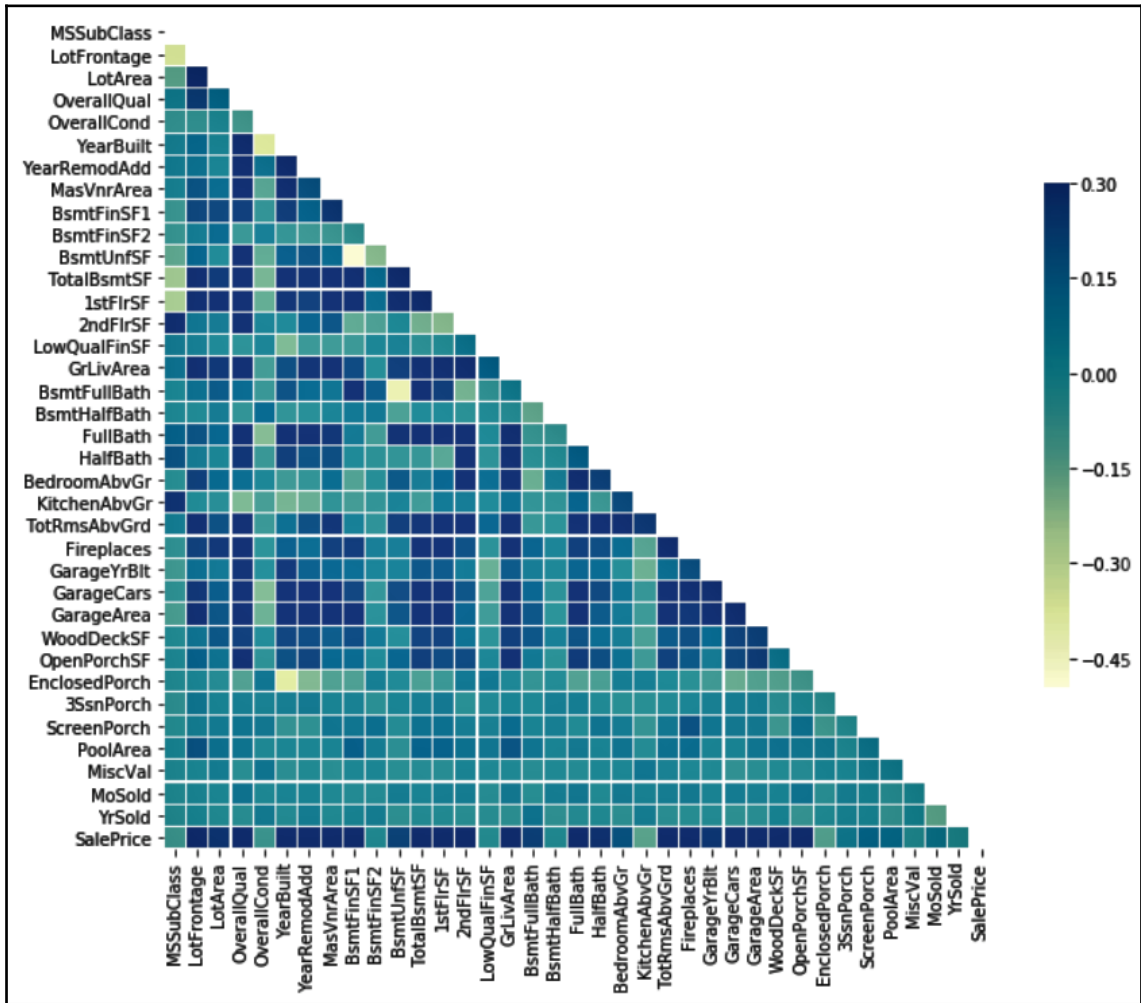


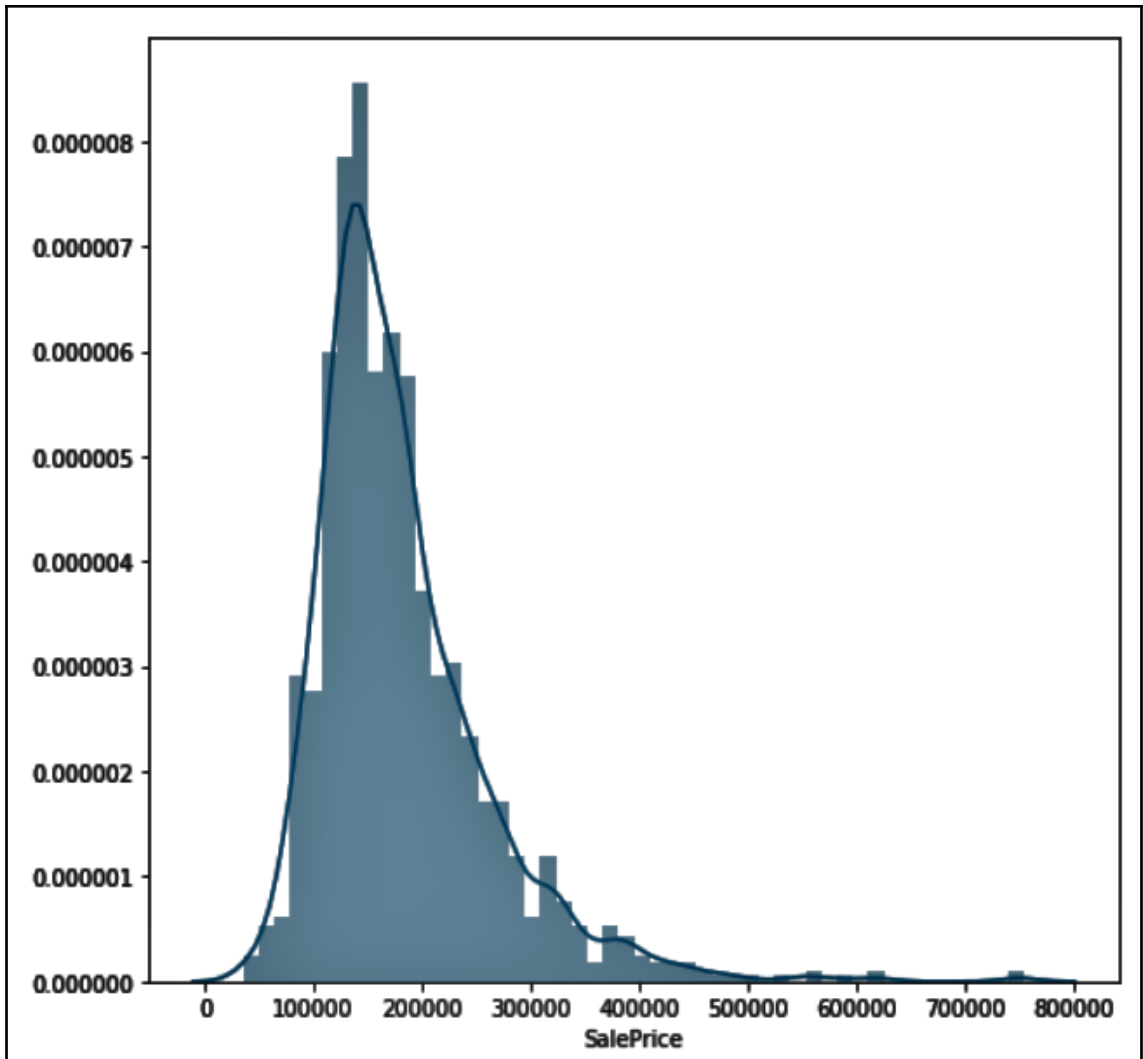


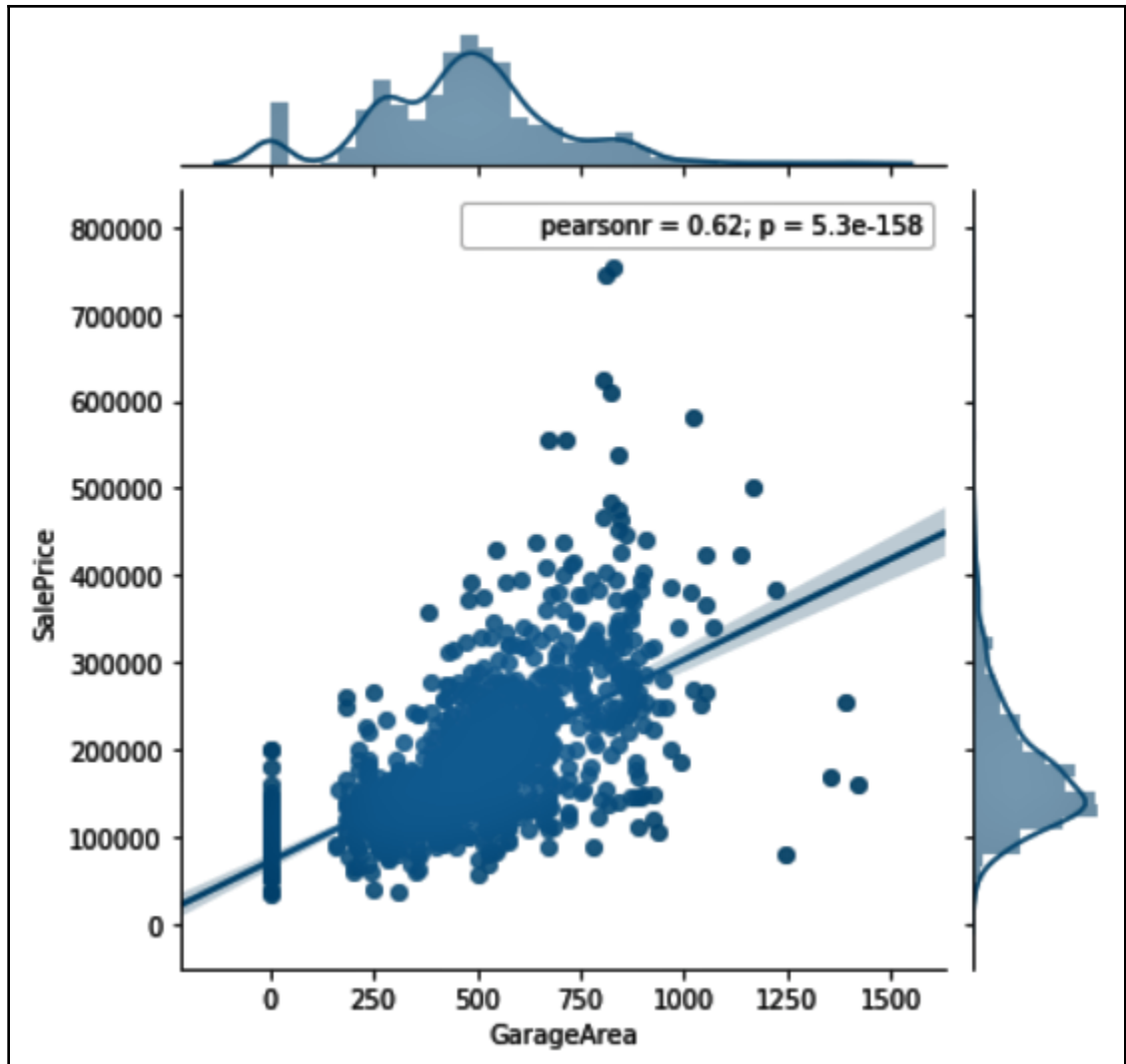




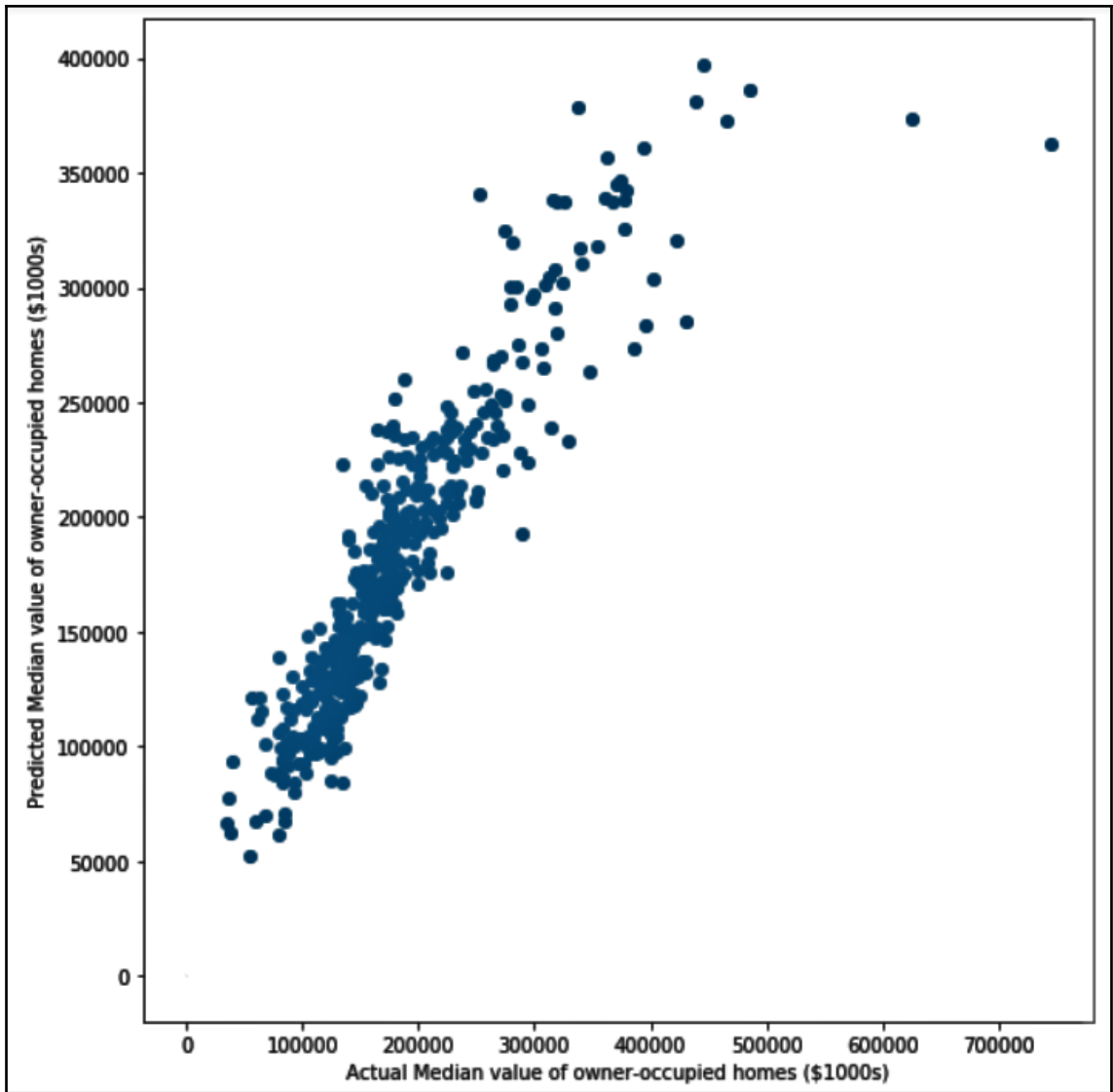
# Chapter 4: Statistical and Machine Learning Algorithms

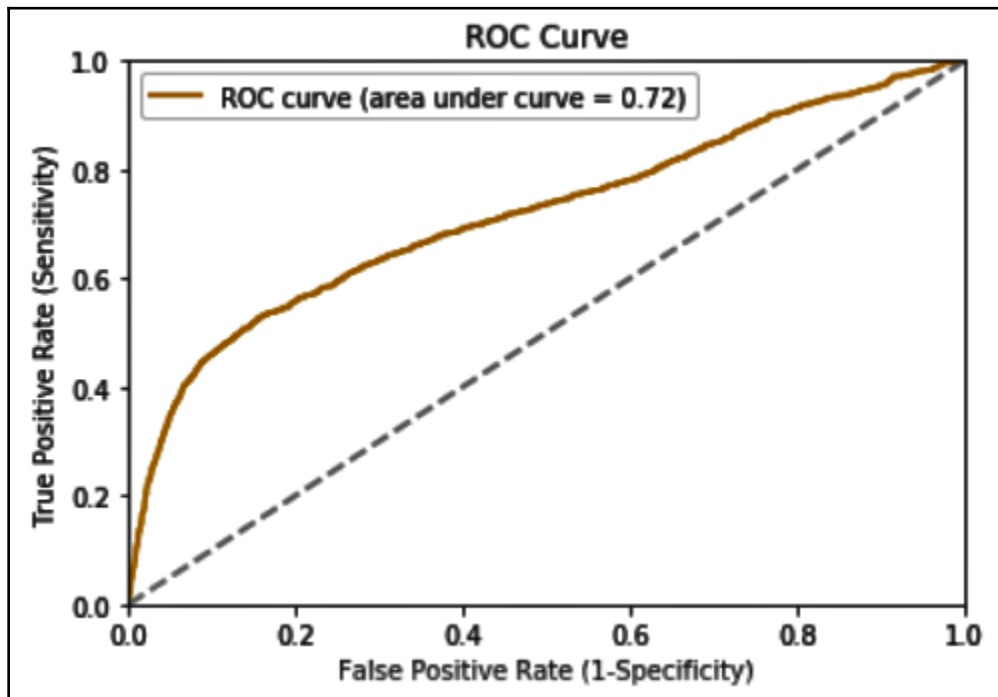
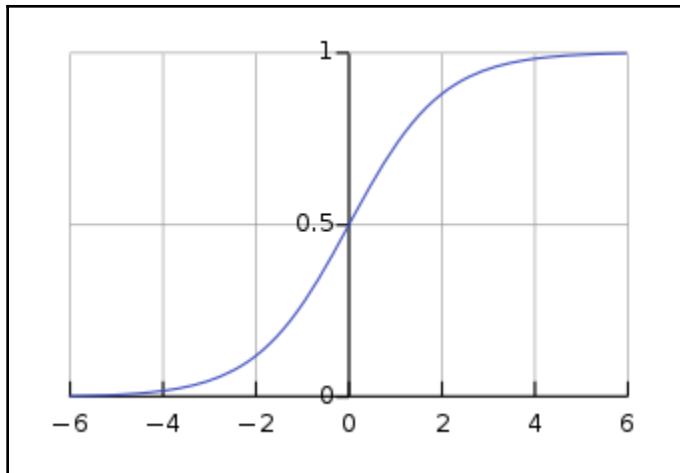






	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1
0	0.235294	0.150685	0.033420	0.666667	0.500	0.949275	0.883333	0.12250	0.125089
1	0.000000	0.202055	0.038795	0.555556	0.875	0.753623	0.433333	0.00000	0.173281
2	0.235294	0.160959	0.046507	0.666667	0.500	0.934783	0.866667	0.10125	0.086109
3	0.294118	0.133562	0.038561	0.666667	0.500	0.311594	0.333333	0.00000	0.038271
4	0.235294	0.215753	0.060576	0.777778	0.500	0.927536	0.833333	0.21875	0.116052





$$p(\textit{class} | \textit{observation}) = \frac{p(\textit{observation} | \textit{class}) \times P(\textit{class})}{p(\textit{observation})}$$

↑ Likelihood
↑ Class Prior Probability

↓ Posterior Probability
↓ Predictor Prior Probability

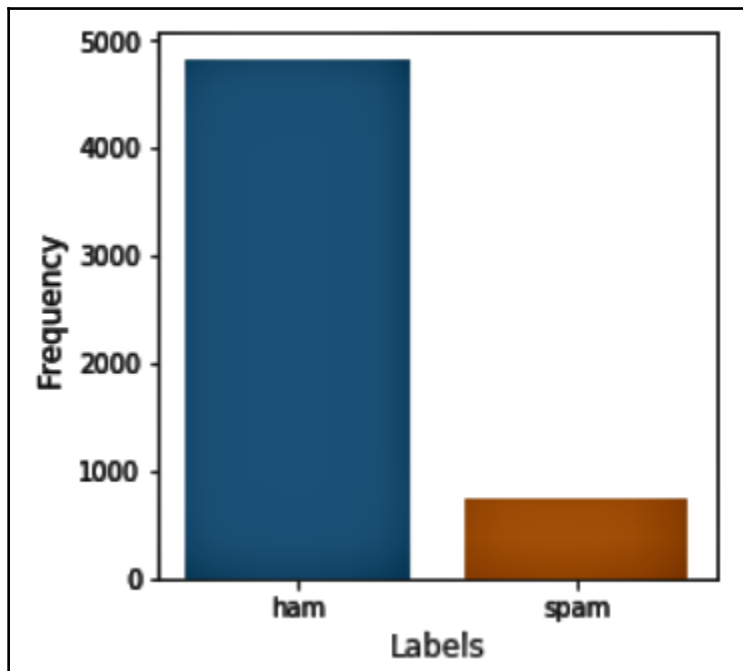
	<b>labels</b>	<b>message</b>
<b>0</b>	ham	Go until jurong point, crazy.. Available only ...
<b>1</b>	ham	Ok lar... Joking wif u oni...
<b>2</b>	spam	Free entry in 2 a wkly comp to win FA Cup fina...

	<b>labels</b>	<b>message</b>
<b>count</b>	5572	5572
<b>unique</b>	2	5169
<b>top</b>	ham	Sorry, I'll call later
<b>freq</b>	4825	30

<b>message</b>				
<b>labels</b>	<b>count</b>	<b>unique</b>	<b>top</b>	<b>freq</b>
<b>ham</b>	4825	4516	Sorry, I'll call later	30
<b>spam</b>	747	653	Please call our customer service representativ...	4

---

	message	word_count	character_count
0	Go until jurong point, crazy.. Available only ...	20	111
1	Ok lar... Joking wif u oni...	6	29
2	Free entry in 2 a wkly comp to win FA Cup fina...	28	155
3	U dun say so early hor... U c already then say...	11	49
4	Nah I don't think he goes to usf, he lives aro...	13	61





	labels	message	word_count	character_count
0	0	Go until jurong point, crazy.. Available only ...	20	111
1	0	Ok lar... Joking wif u oni...	6	29
2	1	Free entry in 2 a wkly comp to win FA Cup fina...	28	155
3	0	U dun say so early hor... U c already then say...	11	49
4	0	Nah I don't think he goes to usf, he lives aro...	13	61

**Accuracy score: 0.9937177473636976**  
**Precision score: 0.9728813559322034**  
**Recall score: 0.9795221843003413**  
**F1 score: 0.976190476190476**

**Accuracy score: 0.9829596412556054**  
**Precision score: 0.961038961038961**  
**Recall score: 0.9192546583850931**  
**F1 score: 0.9396825396825396**

national_inv	0
lead_time	3403
in_transit_qty	0
forecast_3_month	0
forecast_6 month	0

---

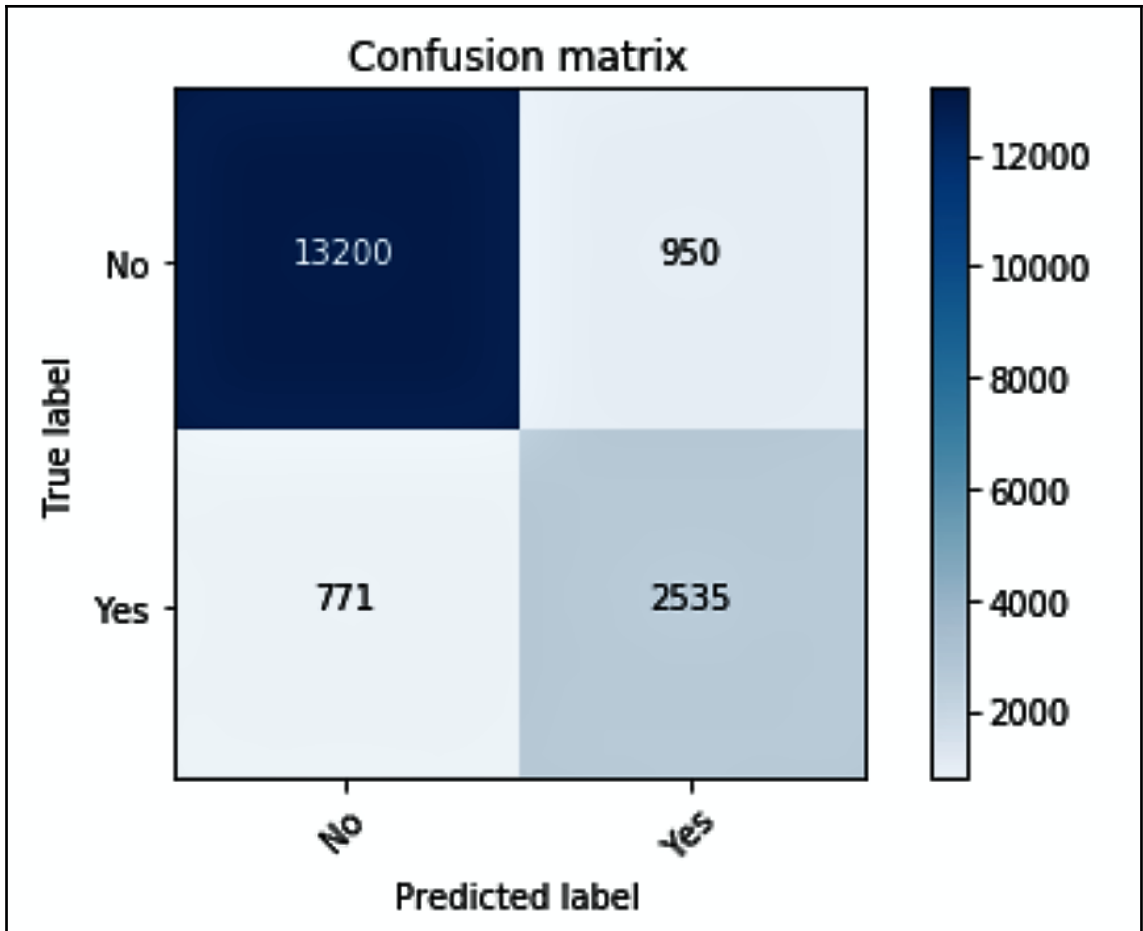
potential_issue_Yes	uint8
deck_risk_Yes	uint8
oe_constraint_Yes	uint8
ppap_risk_Yes	uint8
stop_auto_buy_Yes	uint8
rev_stop_Yes	uint8
went_on_backorder_Yes	uint8

0	47217
1	10969

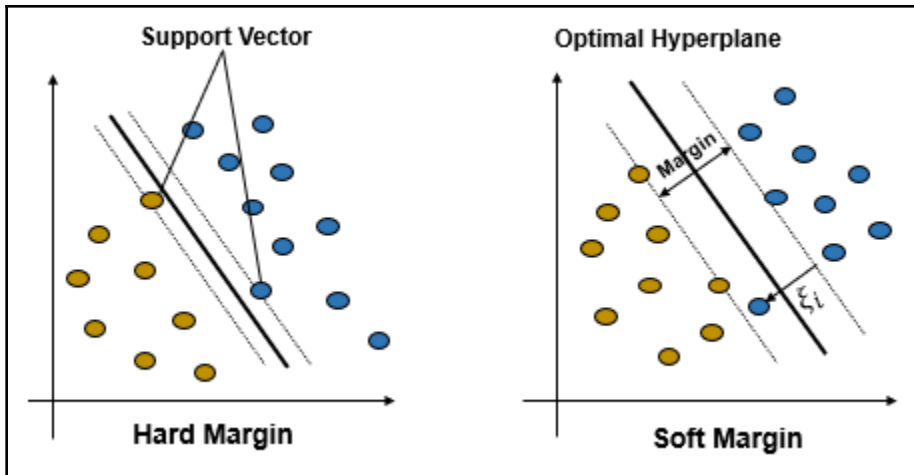
dtype: int64

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, presort=False, random_state=None,
splitter='best')
```

**Accuracy is 0.901409257562**  
**TN: 13200 FP: 950 FN: 771 TP: 2535**



	precision	recall	f1-score	support
No	0.89	0.95	0.92	9439
Yes	0.70	0.48	0.56	2199
avg / total	0.85	0.86	0.85	11638



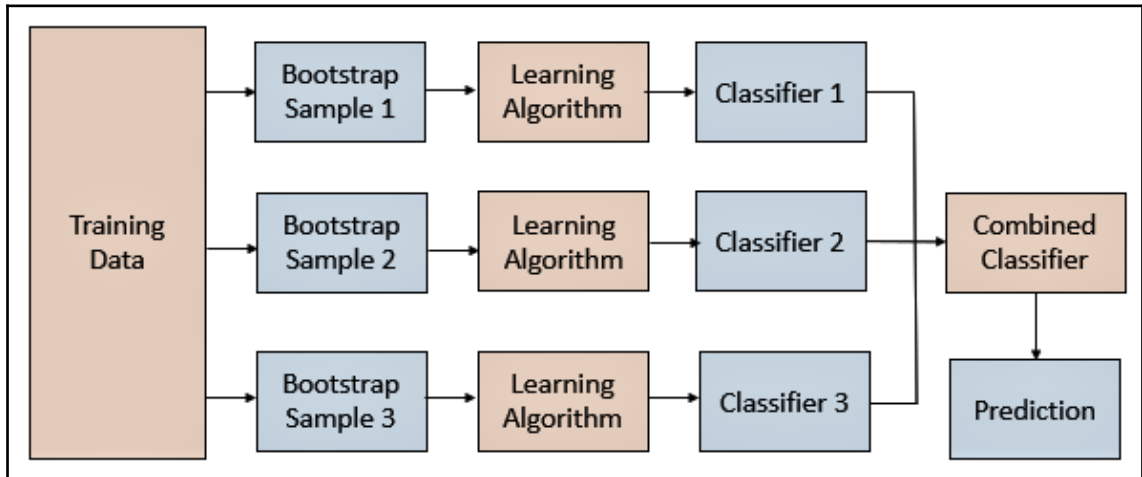
**Train Accuracy Score:**  
 0.99889380531  
**Test Accuracy Score:**  
 0.872928176796

**Train Accuracy Score:**  
 0.998340707965  
**Test Accuracy Score:**  
 0.993370165746

**Train Accuracy Score:**  
 0.997234513274  
**Test Accuracy Score:**  
 0.994475138122

---

## Chapter 5: Bag the Models with Bagging

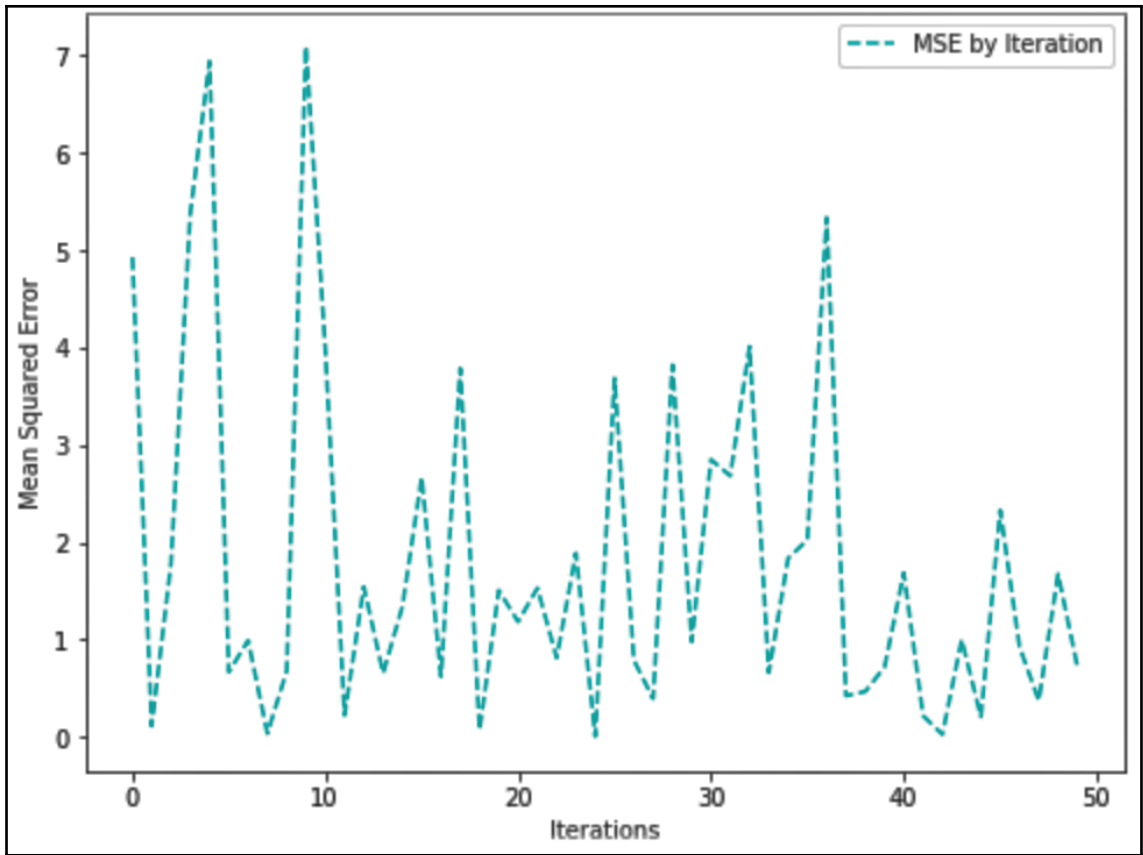


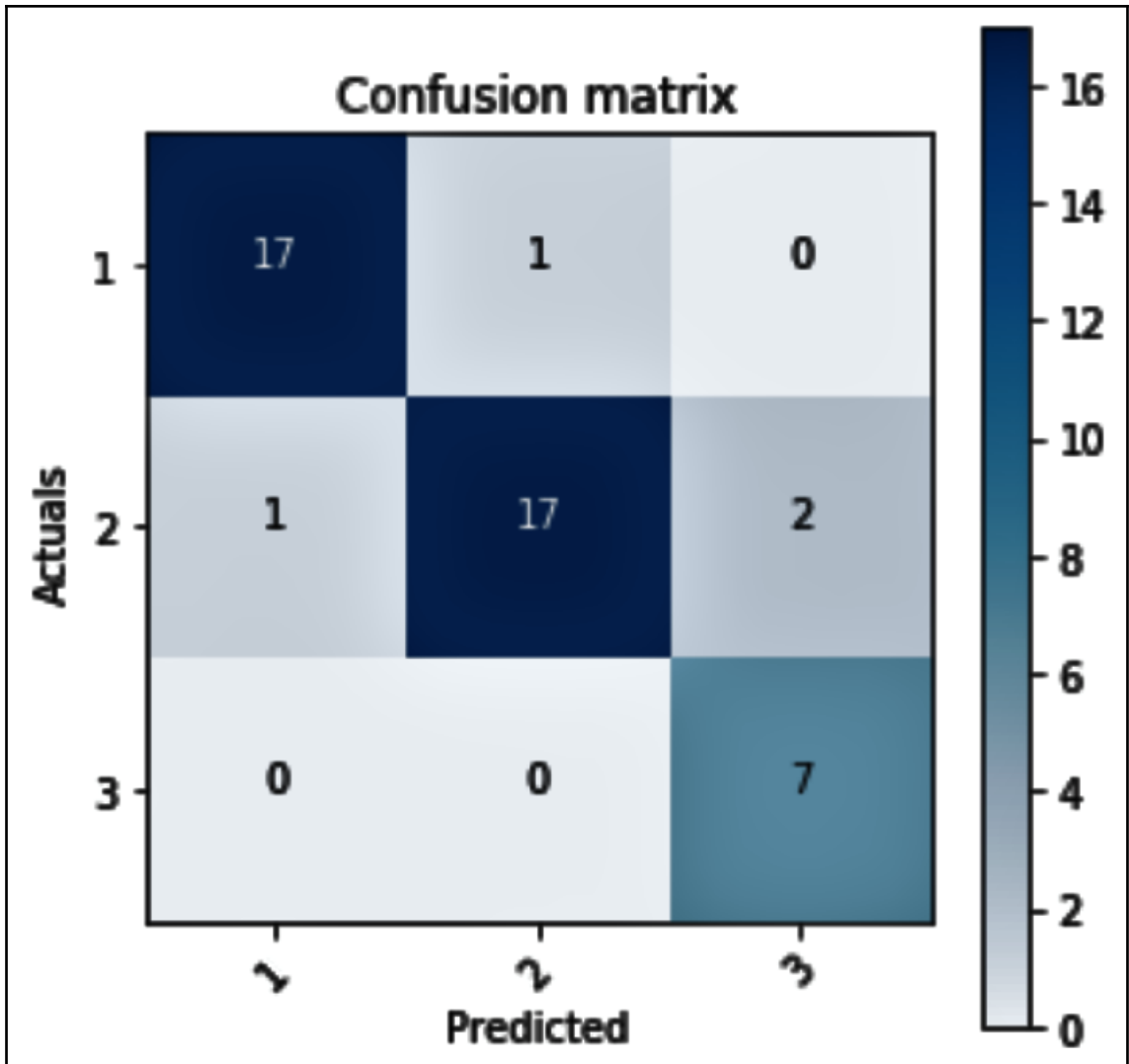
*Bagging*( $S = ((x_1, y_1), \dots, (x_m, y_m))$ )

- 1 for  $t \leftarrow 1$  to  $T$  do
- 2  $S_t \leftarrow \text{Bootstrap}(S) \triangleright$  *i.i.d. Sampling with replacement from  $S$ .*
- 3  $h_t \leftarrow \text{TrainClassifier}(S_t)$
- 4 return  $h_s = x \rightarrow \text{MajorityVote}((h_1(x), \dots, (h_T(x)))$ )

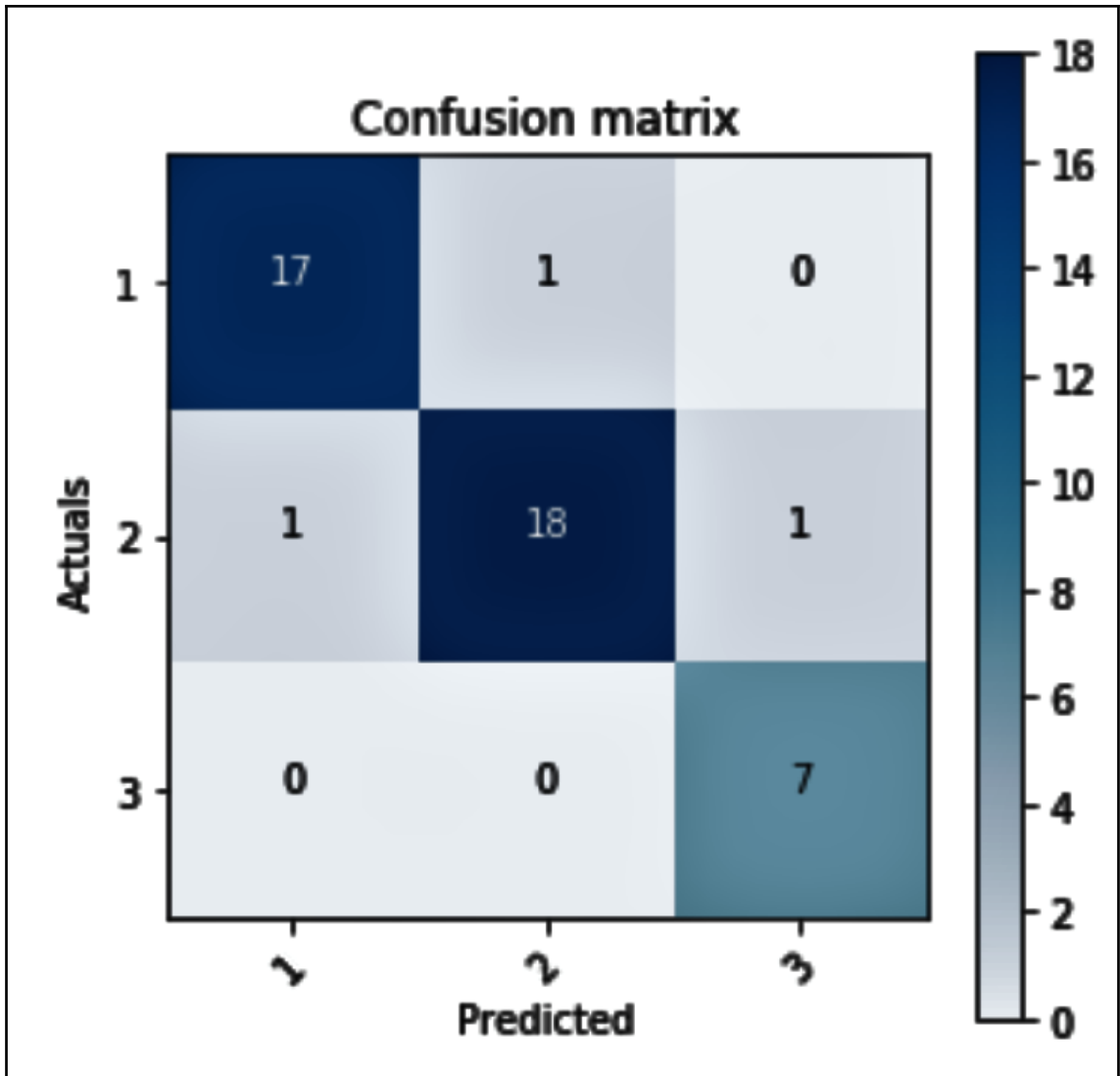
*Bagging*( $S = ((x_1, y_1), \dots, (x_m, y_m))$ )

- 1 for  $t \leftarrow 1$  to  $T$  do
- 2  $S_t \leftarrow \text{Bootstrap}(S) \triangleright$  *i.i.d. Sampling with replacement from  $S$ .*
- 3  $h_t \leftarrow \text{TrainRegression}(S_t)$
- 4 return  $h_s = x \rightarrow \text{Mean}((h_1(x), \dots, (h_T(x)))$ )

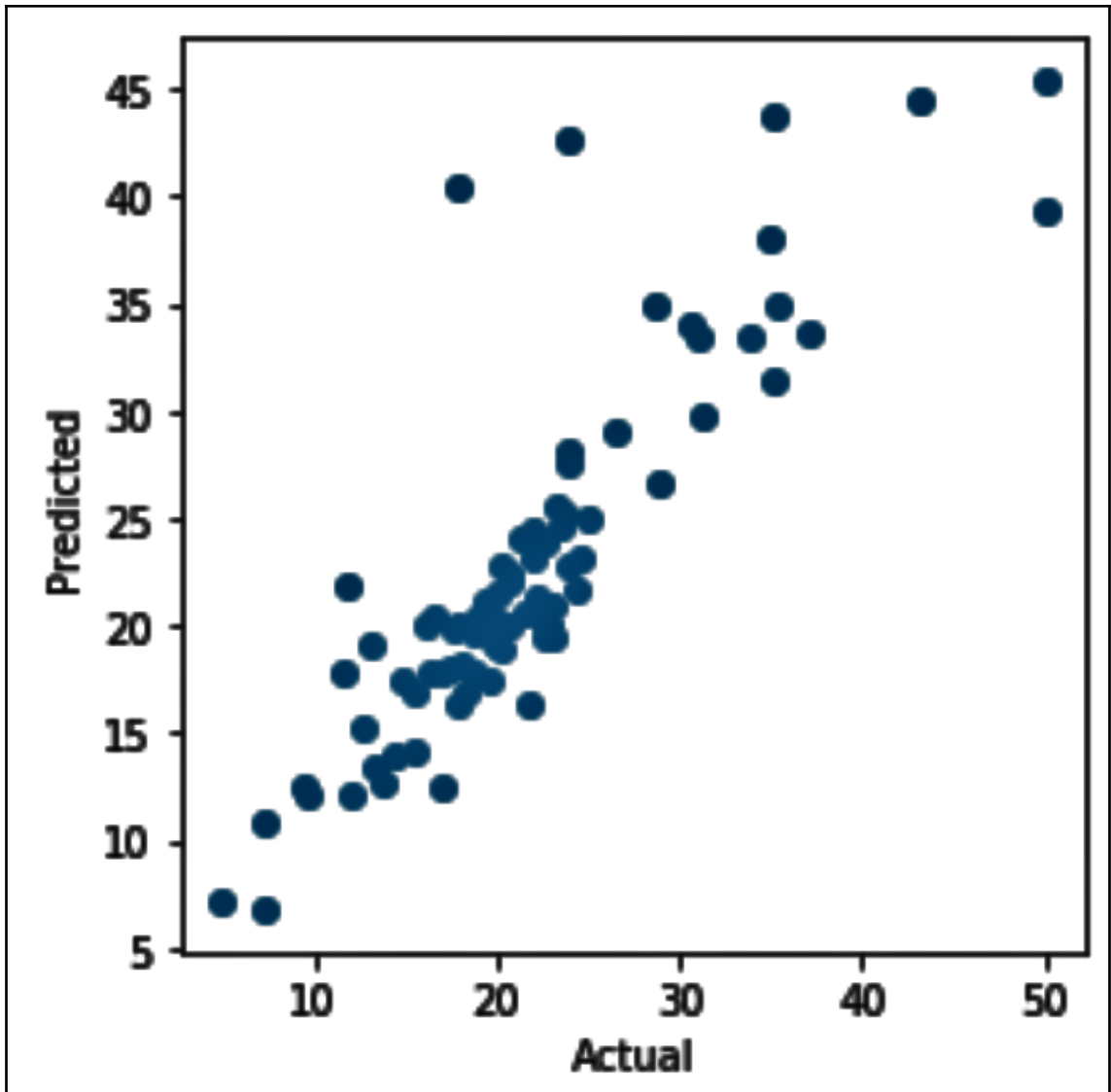




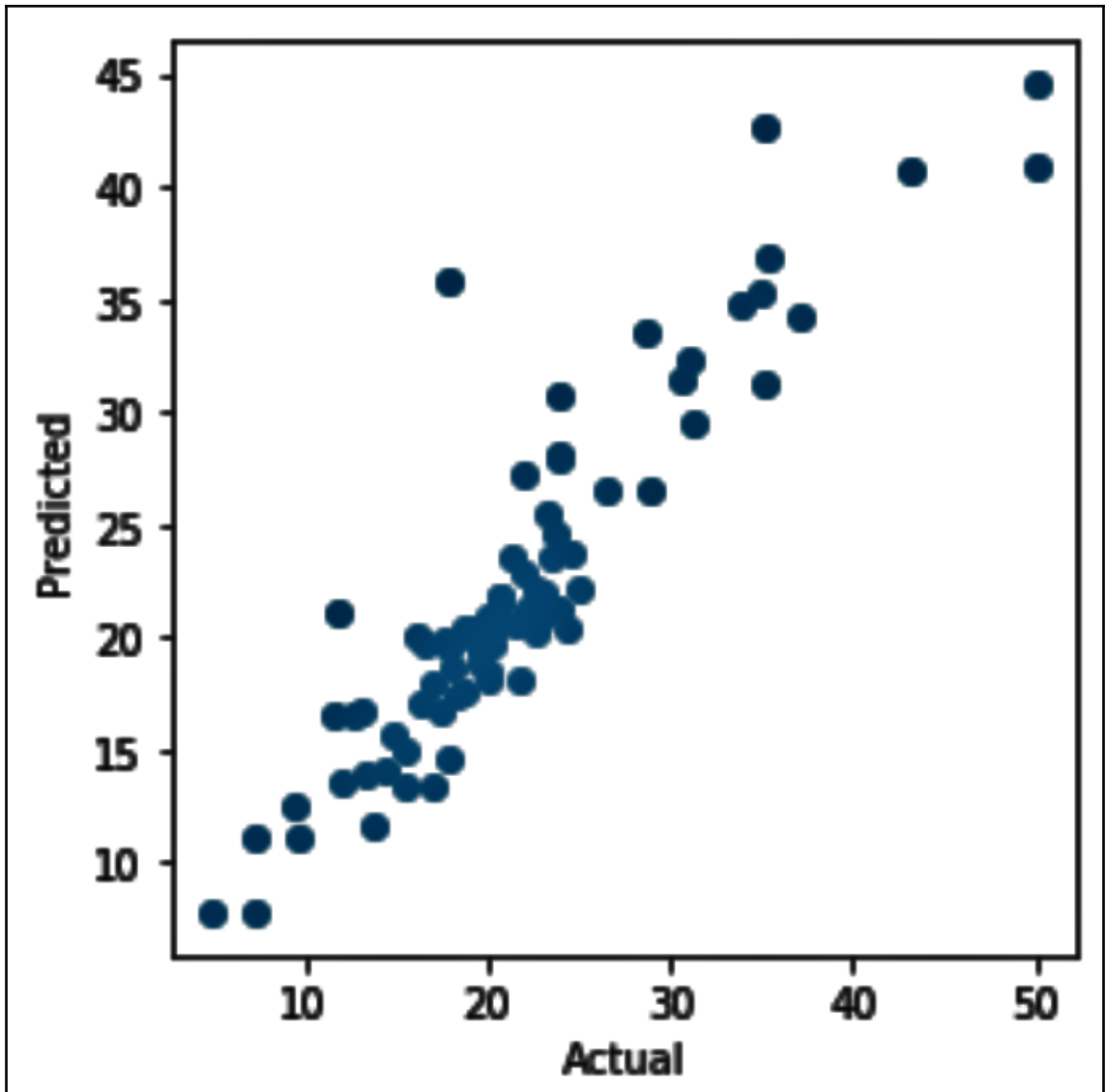
```
{'base_estimator__max_depth': 3, 'base_estimator__max_leaf_nodes': 10, 'n_estimators': 20}
```



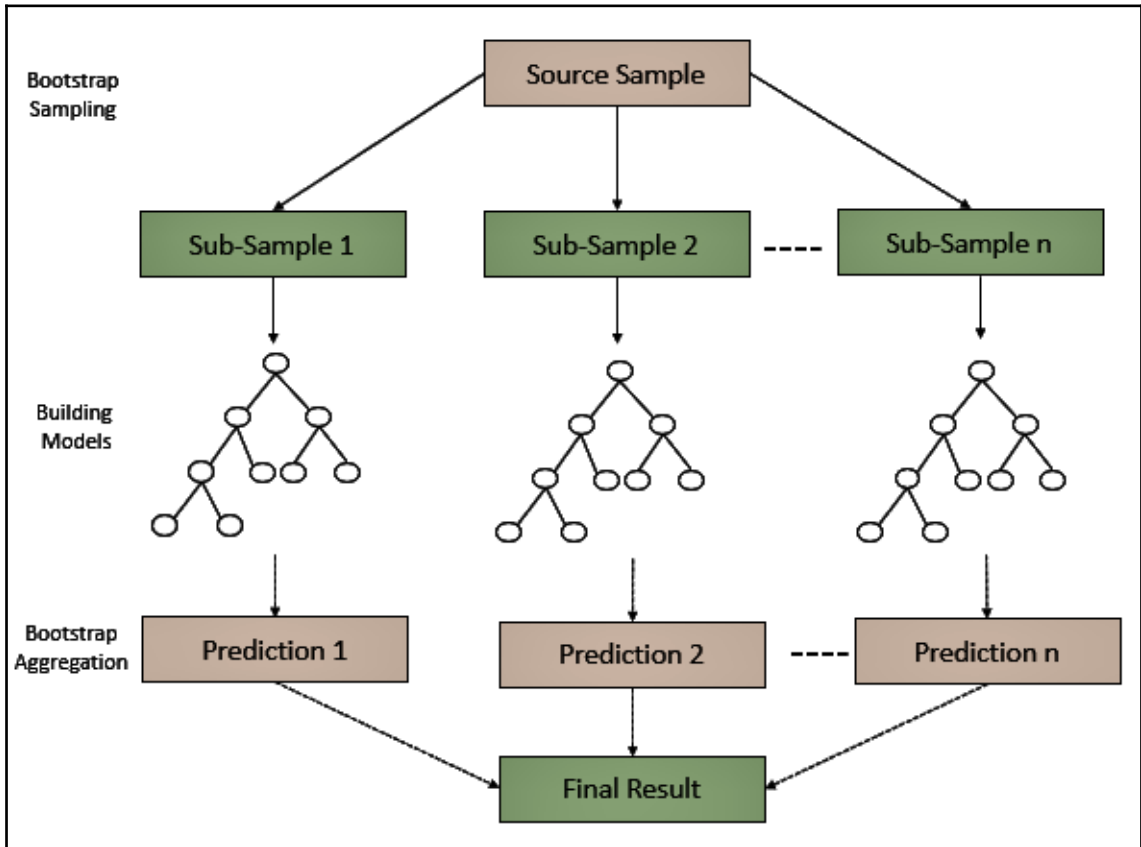


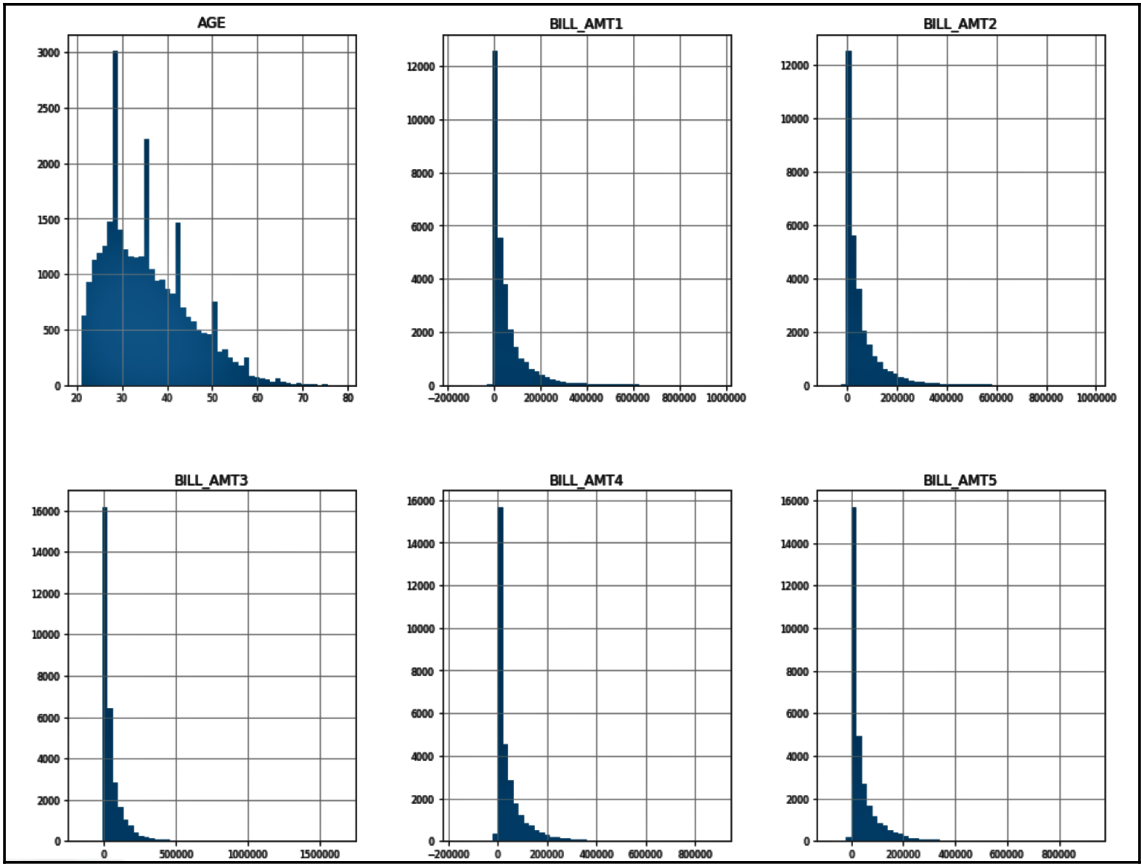


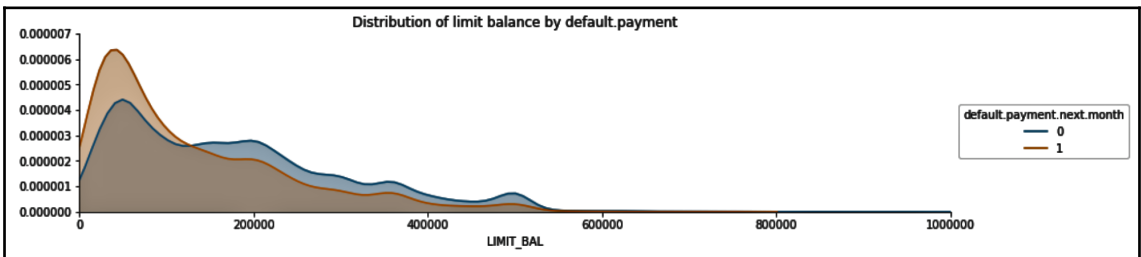
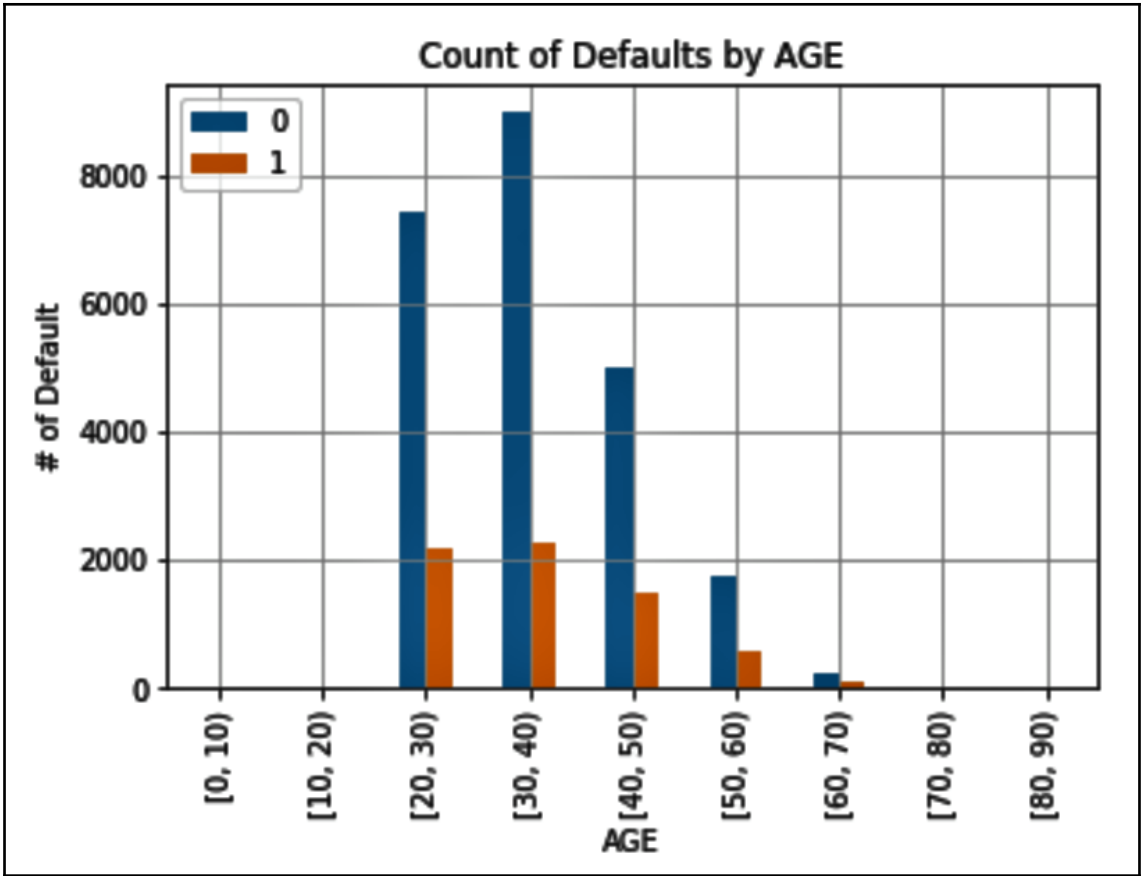
0.82224534363352664



# Chapter 6: When in Doubt, Use Random Forests





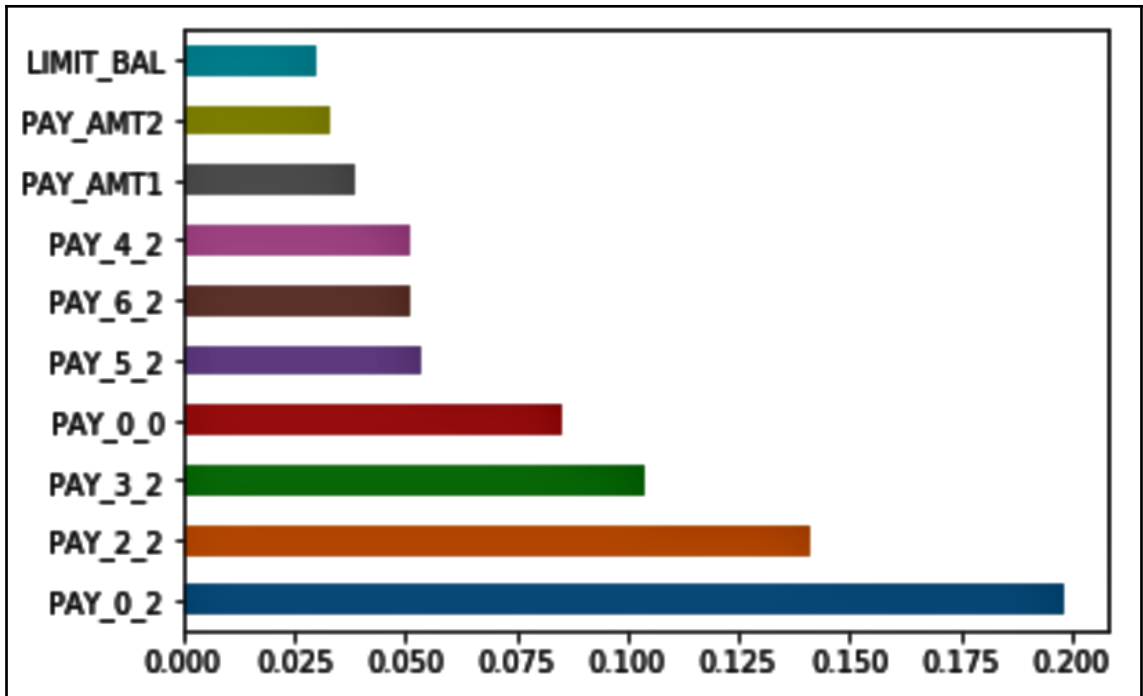


---

```
Accuracy Score          0.811889
Kappa Score             0.278229
Model                   Random Forest Classifier
Precision Score         0.673212
ROC Score               0.60721
Recall Score            0.247423
dtype: object
```

```
Accuracy Score          0.811889
Kappa Score             0.278229
Model                   Random Forest Classifier
Precision Score         0.673212
ROC Score               0.60721
Recall Score            0.247423
dtype: object
```

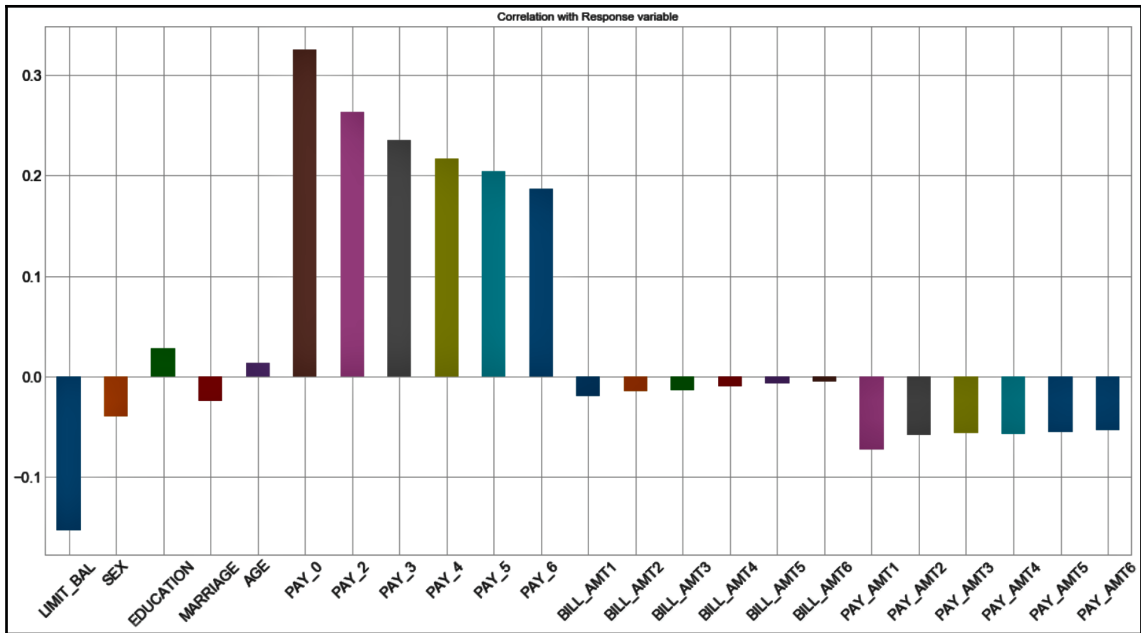
	precision	recall	f1-score	support
0	0.82	0.97	0.89	7060
1	0.67	0.25	0.36	1940
avg / total	0.79	0.81	0.78	9000



---

```
Connecting to H2O server at http://127.0.0.1:54321... successful.  
H2O cluster uptime:          02 secs  
H2O cluster timezone:       Etc/UTC  
H2O data parsing timezone:   UTC  
H2O cluster version:        3.22.1.2  
H2O cluster version age:     4 hours and 18 minutes  
H2O cluster name:           H2O_from_python_unknownUser_zhwtlv  
H2O cluster total nodes:    1  
H2O cluster free memory:     2.938 Gb  
H2O cluster total cores:    2  
H2O cluster allowed cores:  2  
H2O cluster status:         accepting new members, healthy  
H2O connection url:         http://127.0.0.1:54321  
H2O connection proxy:       None  
H2O internal security:      False  
H2O API Extensions:         XGBoost, Algos, AutoML, Core V3, Core V4  
Python version:             3.6.7 final
```





ModelMetricsBinomial: drf  
 \*\* Reported on test data. \*\*

MSE: 0.13916702510839007

RMSE: 0.3730509685128697

LogLoss: 0.4404003566729752

Mean Per-Class Error: 0.3021721562099693

AUC: 0.7689912083007103

pr\_auc: 0.5367690443947648

Gini: 0.5379824166014207

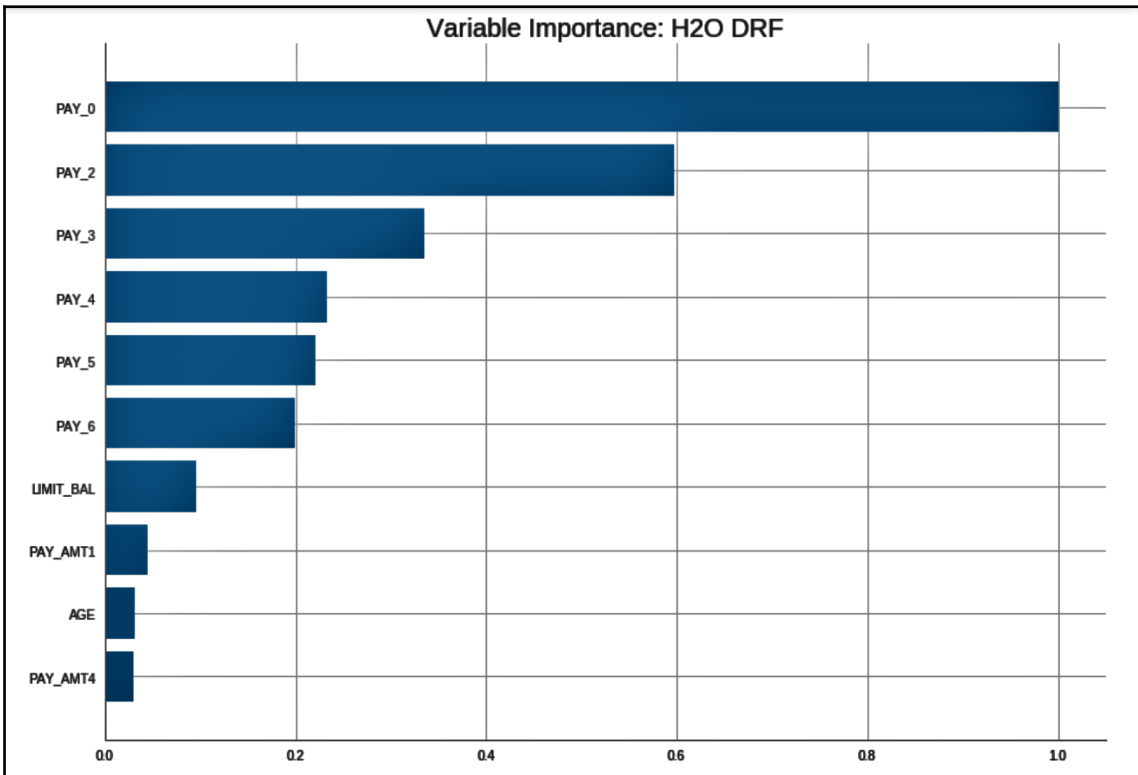
Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.31553293049335485:

	0	1	Error	Rate
0	6035.0	902.0	0.13	(902.0/6937.0)
1	969.0	1034.0	0.4838	(969.0/2003.0)
Total	7004.0	1936.0	0.2093	(1871.0/8940.0)

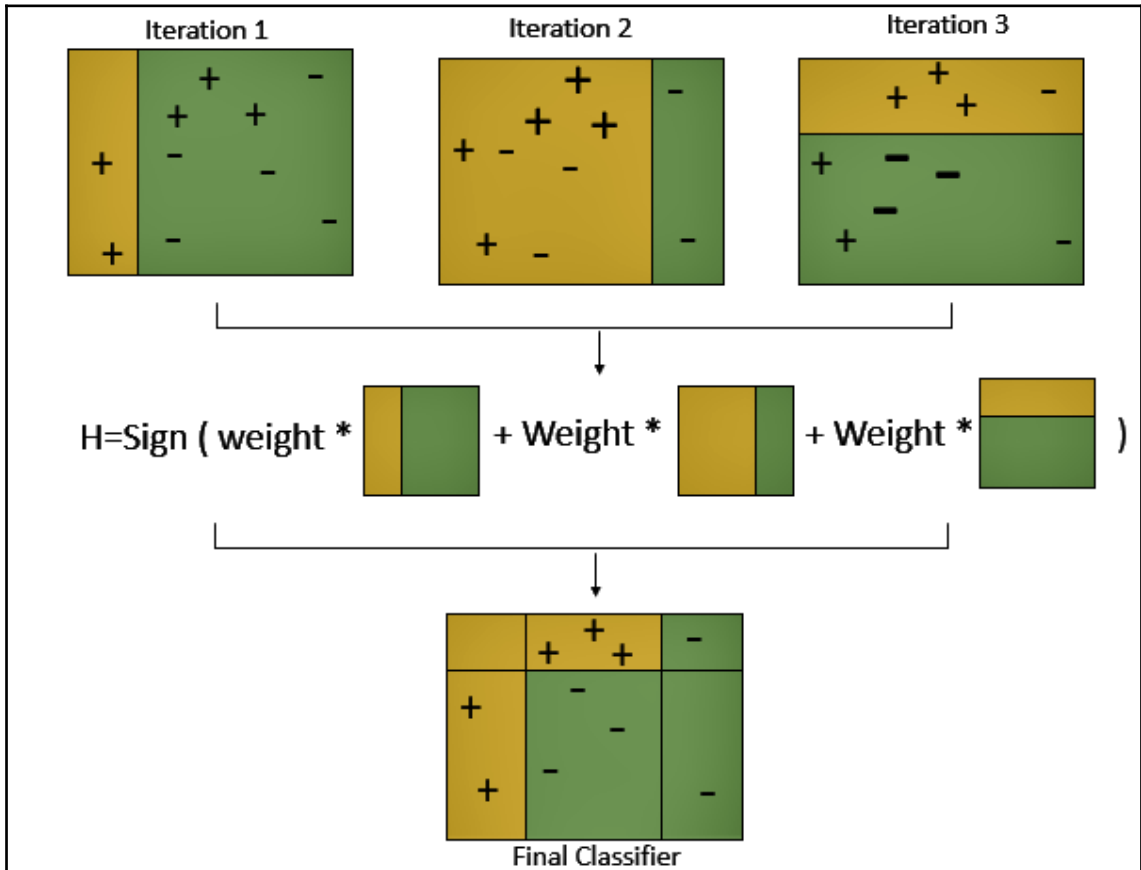
ModelMetricsBinomial: drf  
\*\* Reported on test data. \*\*

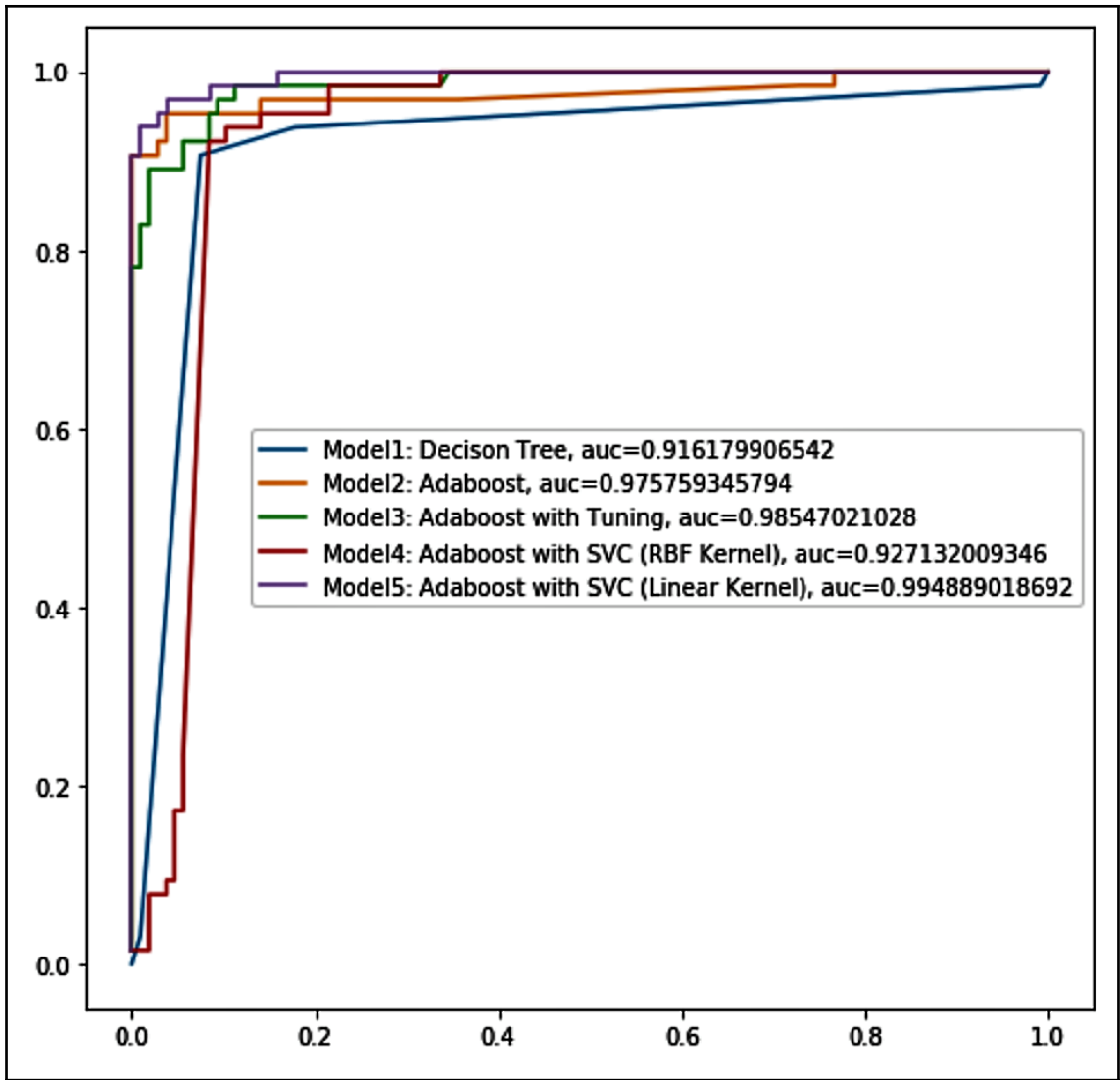
MSE: 0.1377514905065363  
RMSE: 0.3711488791664826  
LogLoss: 0.4364918401503112  
Mean Per-Class Error: 0.29563198088840503  
AUC: 0.7738851575598977  
pr\_auc: 0.5456525571043274  
Gini: 0.5477703151197955  
Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.300957882463187:

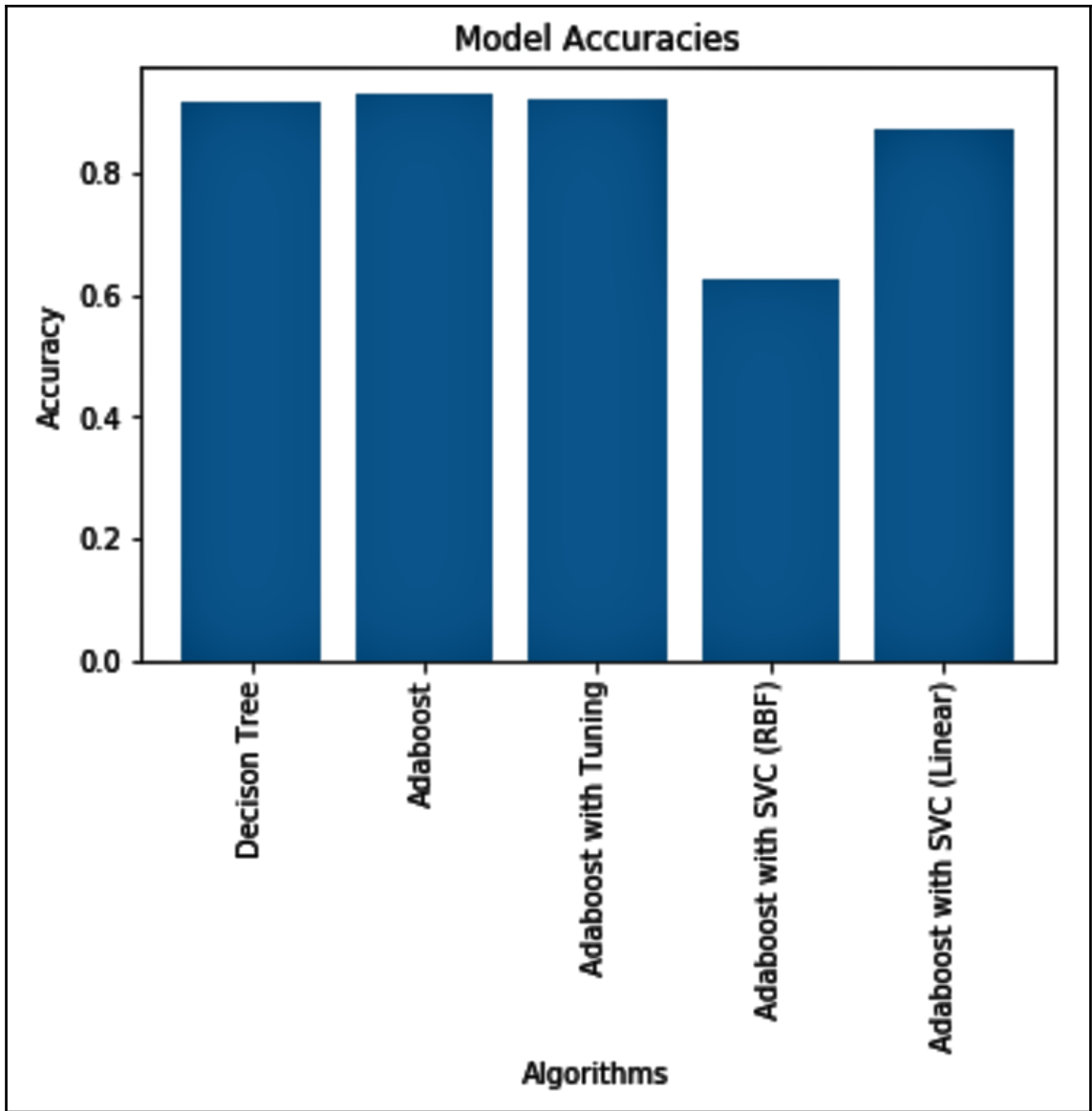
	0	1	Error	Rate
0	6027.0	910.0	0.1312	(910.0/6937.0)
1	945.0	1058.0	0.4718	(945.0/2003.0)
Total	6972.0	1968.0	0.2075	(1855.0/8940.0)



# Chapter 7: Boosting Model Performance with Boosting







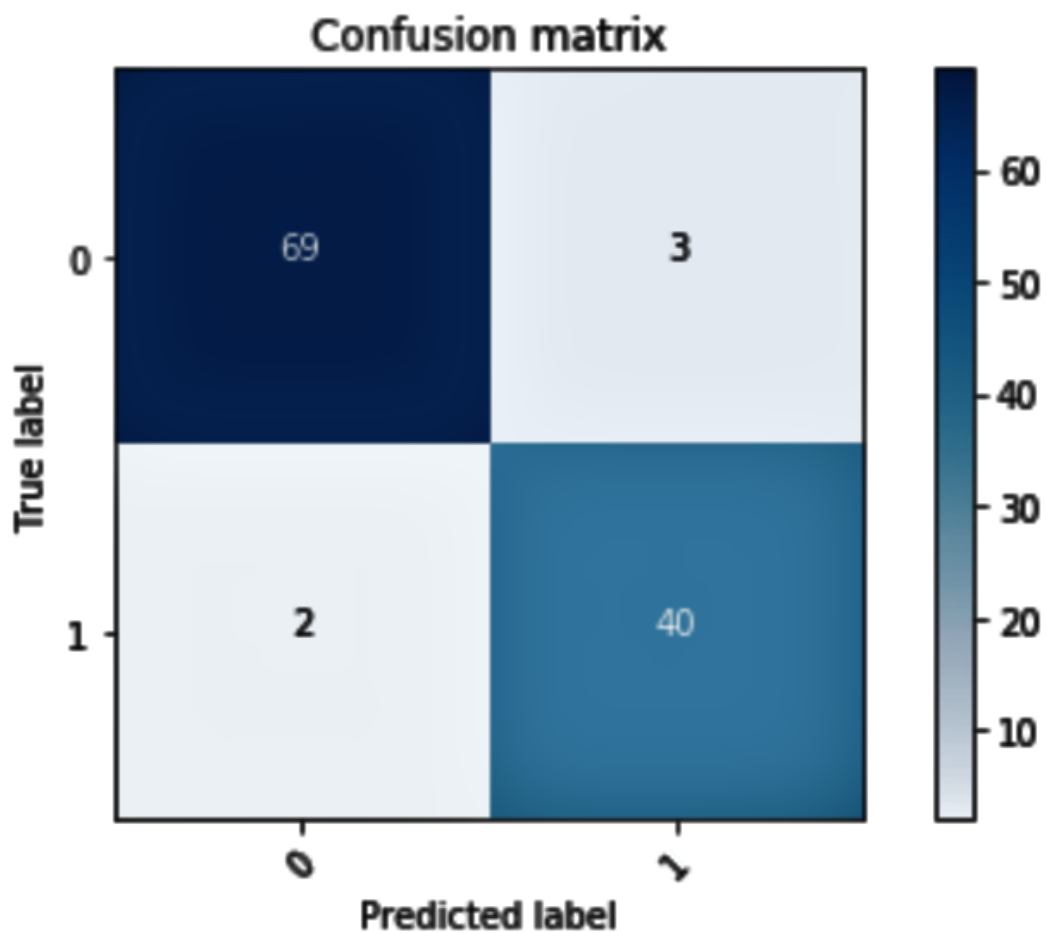
---

	precision	recall	f1-score	support
0	0.97	0.96	0.97	72
1	0.93	0.95	0.94	42
avg / total	0.96	0.96	0.96	114

---

Confusion matrix

```
[[69  3]
 [ 2 40]]
```



---

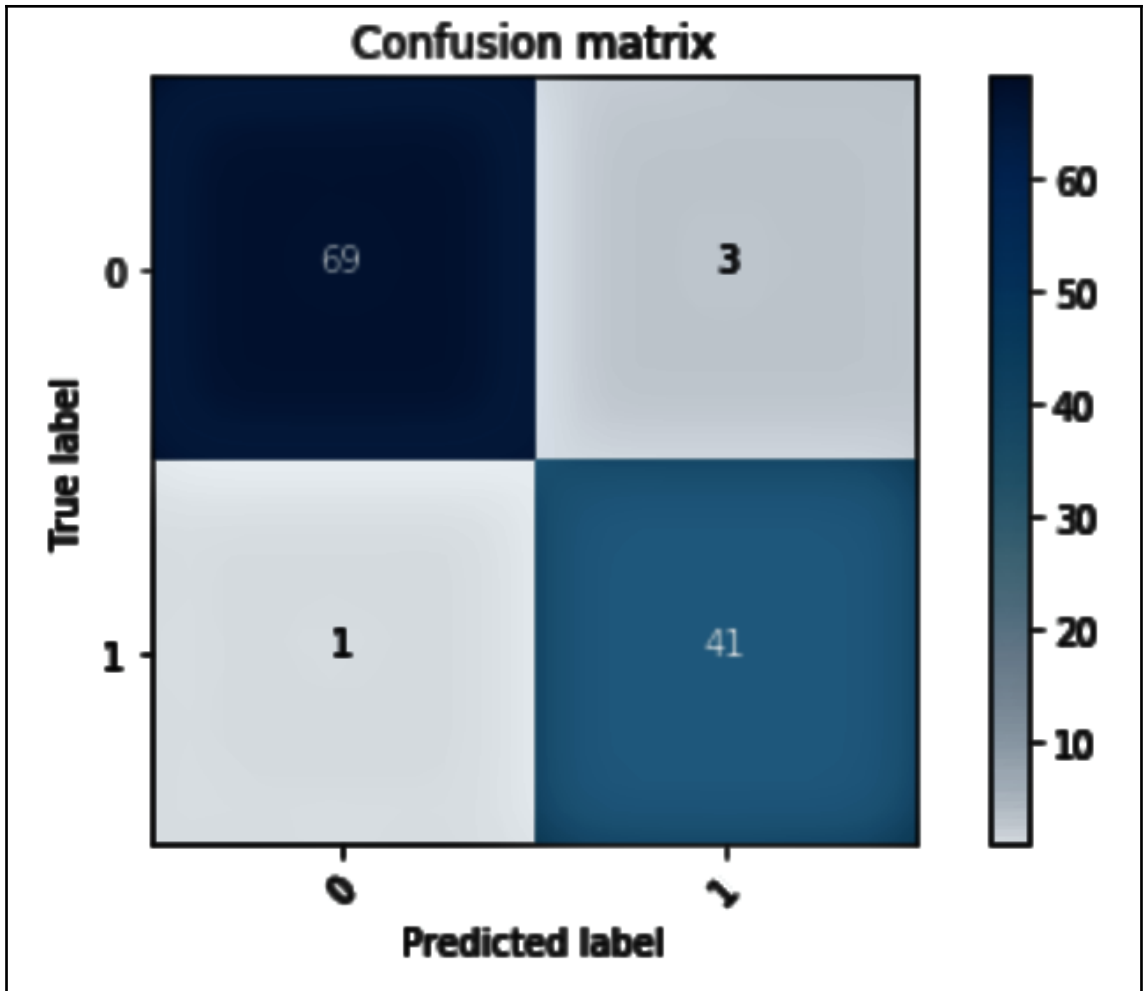
Accuracy score = 0.96

Area under ROC curve = 0.96

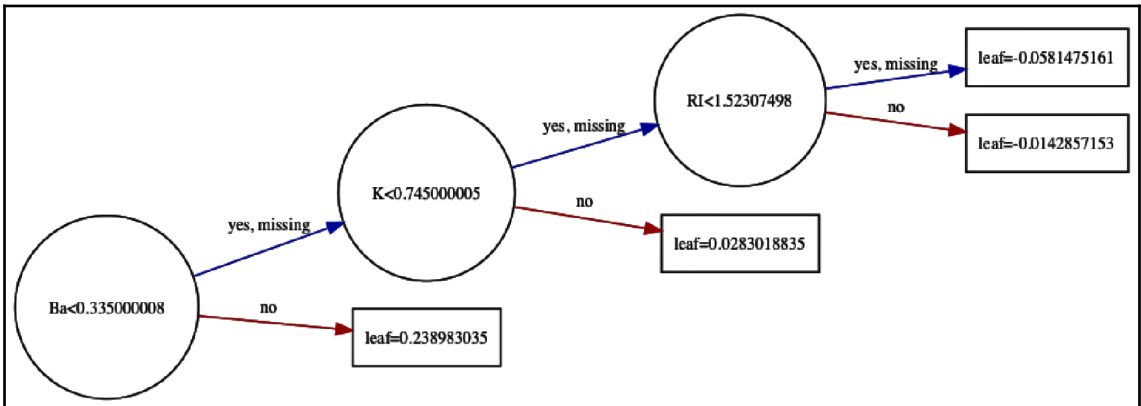
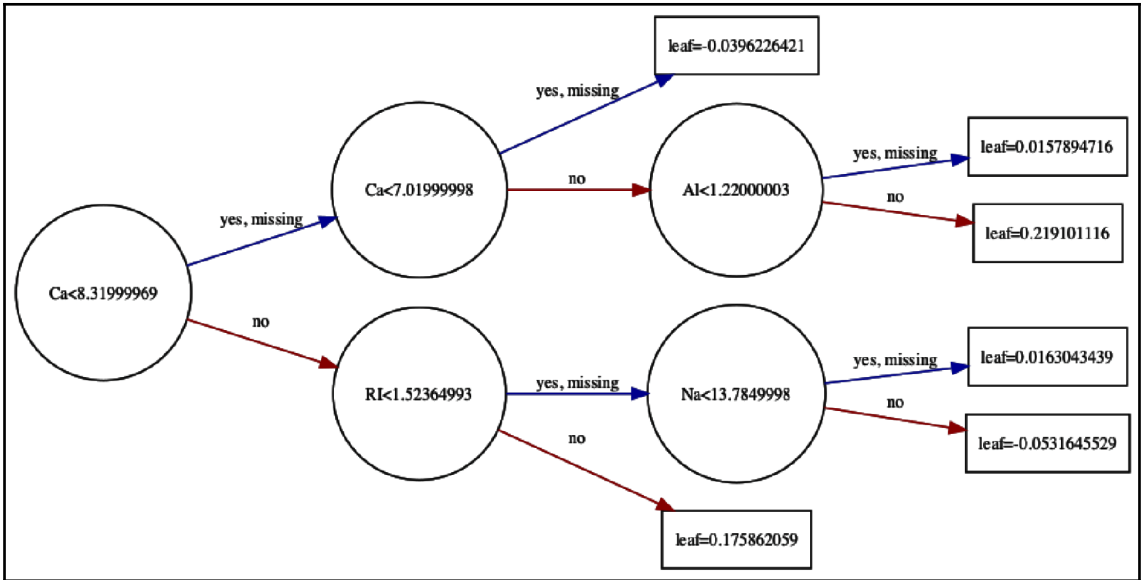
```
GradientBoostingClassifier(criterion='friedman_mse', init=None,
                           learning_rate=0.6, loss='deviance', max_depth=3,
                           max_features='sqrt', max_leaf_nodes=None,
                           min_impurity_decrease=0.0, min_impurity_split=None,
                           min_samples_leaf=0.10000000000000001,
                           min_samples_split=0.10000000000000001,
                           min_weight_fraction_leaf=0.0, n_estimators=100,
                           presort='auto', random_state=None, subsample=0.6, verbose=0,
                           warm_start=False)
```

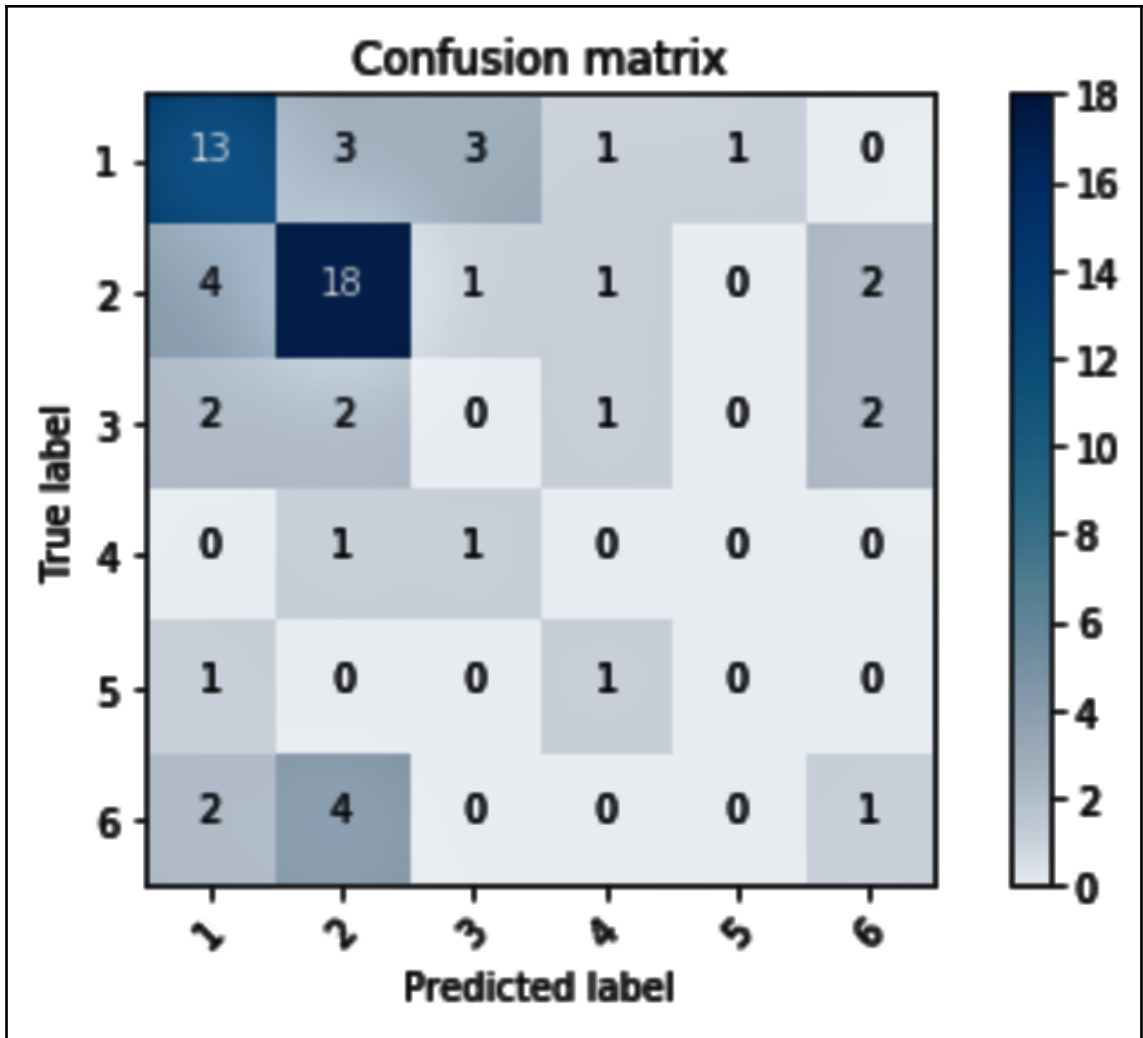
	precision	recall	f1-score	support
0	0.99	0.96	0.97	72
1	0.93	0.98	0.95	42
avg / total	0.97	0.96	0.97	114

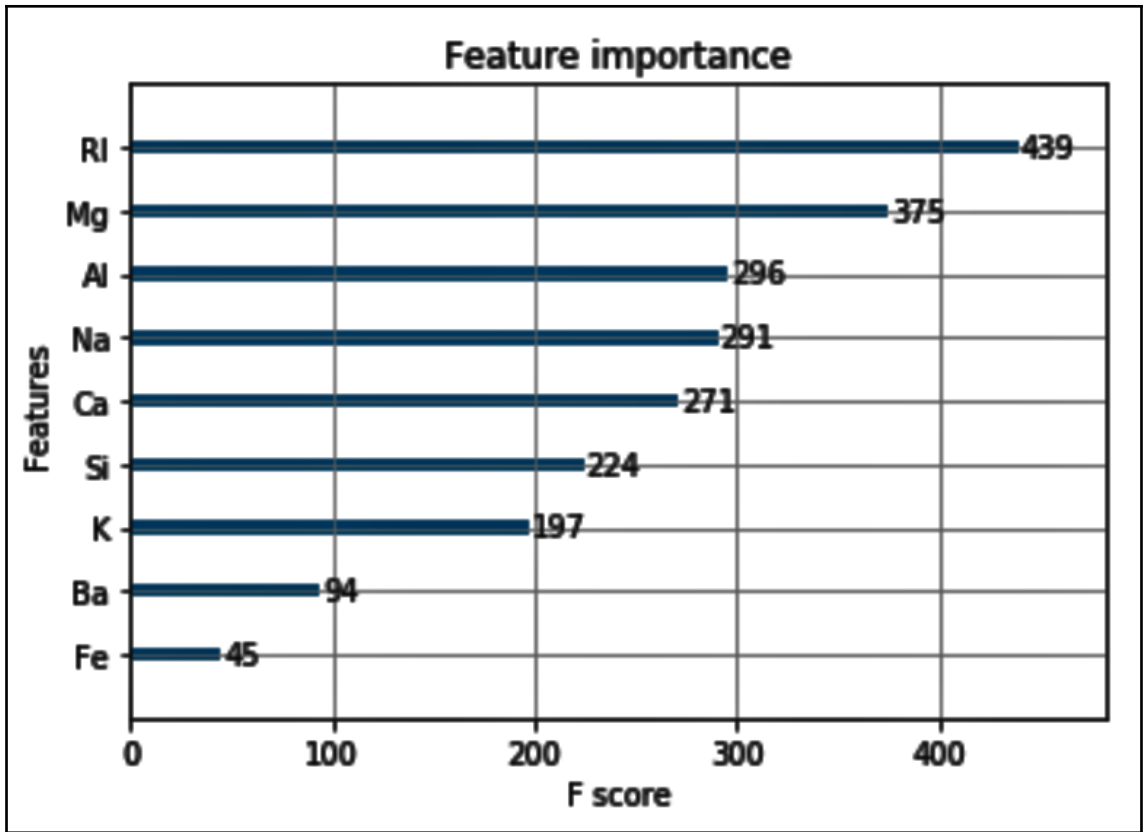




Accuracy score = 0.96  
Area under ROC curve = 0.97





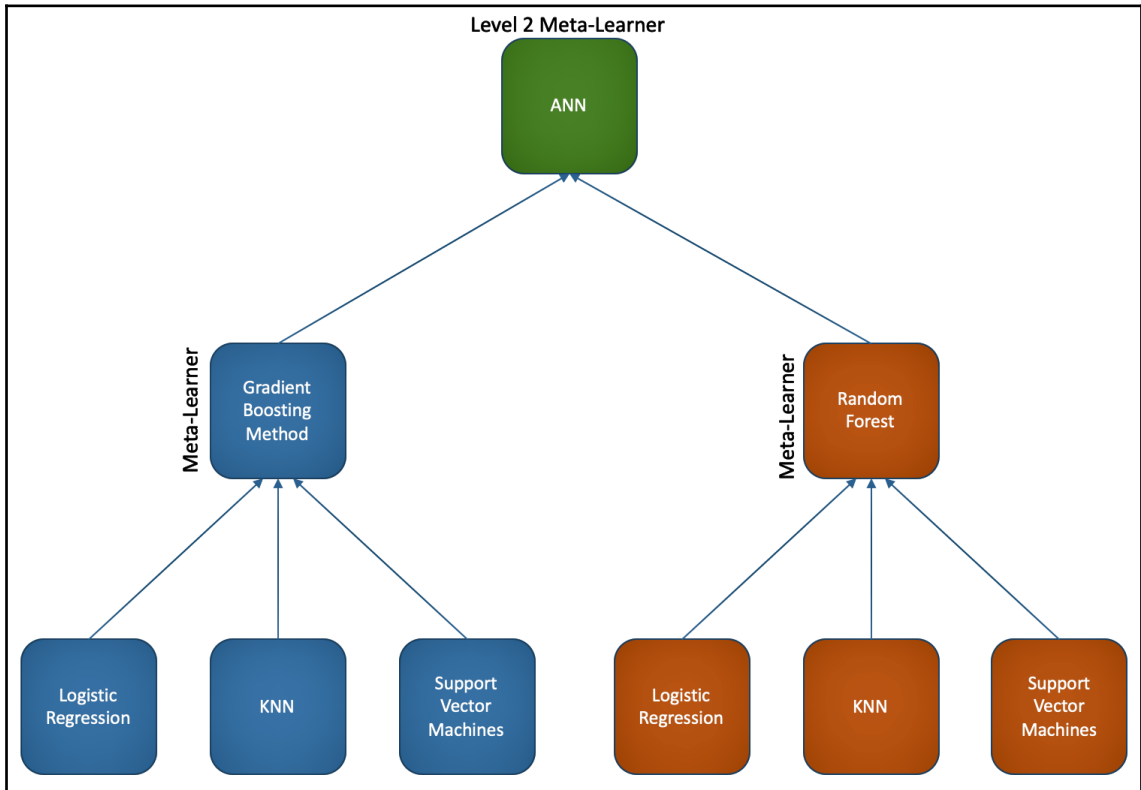


```

Threshold=0.02016, Number of Features=9, Model Accuracy: 0.69%
Threshold=0.04211, Number of Features=8, Model Accuracy: 0.71%
Threshold=0.08826, Number of Features=7, Model Accuracy: 0.69%
Threshold=0.10036, Number of Features=6, Model Accuracy: 0.68%
Threshold=0.12142, Number of Features=5, Model Accuracy: 0.72%
Threshold=0.13038, Number of Features=4, Model Accuracy: 0.71%
Threshold=0.13262, Number of Features=3, Model Accuracy: 0.66%
Threshold=0.16801, Number of Features=2, Model Accuracy: 0.69%
Threshold=0.19668, Number of Features=1, Model Accuracy: 0.49%

```

## Chapter 8: Blend It with Stacking



(5400, 4)				
	NB_VAL	KNN_VAL	DT_VAL	Y_VAL
0	1	0	0	0
1	1	0	0	1
2	1	0	0	0
3	1	0	1	1
4	1	0	0	0

---

(3000, 3)			
	NB_TEST	KNN_TEST	DT_TEST
0	1	0	0
1	1	1	0
2	0	1	1
3	1	1	1
4	1	0	1

Accuracy from GaussianNB: 0.39133333333333
Accuracy from KNN: 0.697
Accuracy from Decision Tree: 0.7326666666667

Accuracy from Meta Learner: 0.7746666666667
---

---

Connecting to H2O server at <http://127.0.0.1:54321>... successful.

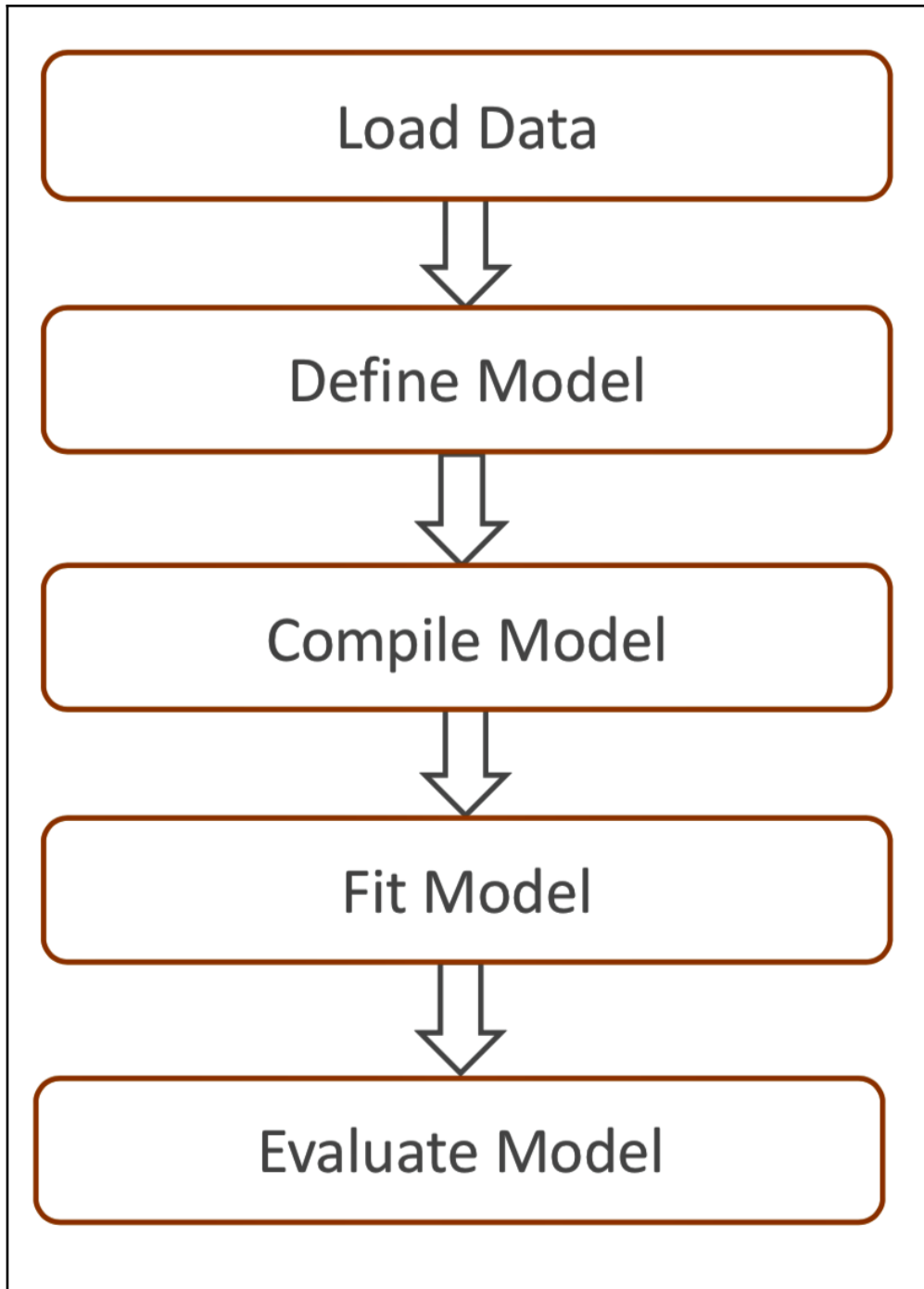
H2O cluster uptime:	02 secs
H2O cluster timezone:	Asia/Kolkata
H2O data parsing timezone:	UTC
H2O cluster version:	3.20.0.8
H2O cluster version age:	3 months and 2 days
H2O cluster name:	H2O_from_python_Dippies_s1thwh
H2O cluster total nodes:	1
H2O cluster free memory:	3.556 Gb
H2O cluster total cores:	8
H2O cluster allowed cores:	8
H2O cluster status:	accepting new members, healthy
H2O connection url:	<a href="http://127.0.0.1:54321">http://127.0.0.1:54321</a>
H2O connection proxy:	None
H2O internal security:	False
H2O API Extensions:	XGBoost, Algos, AutoML, Core V3, Core V4
Python version:	3.6.5 final

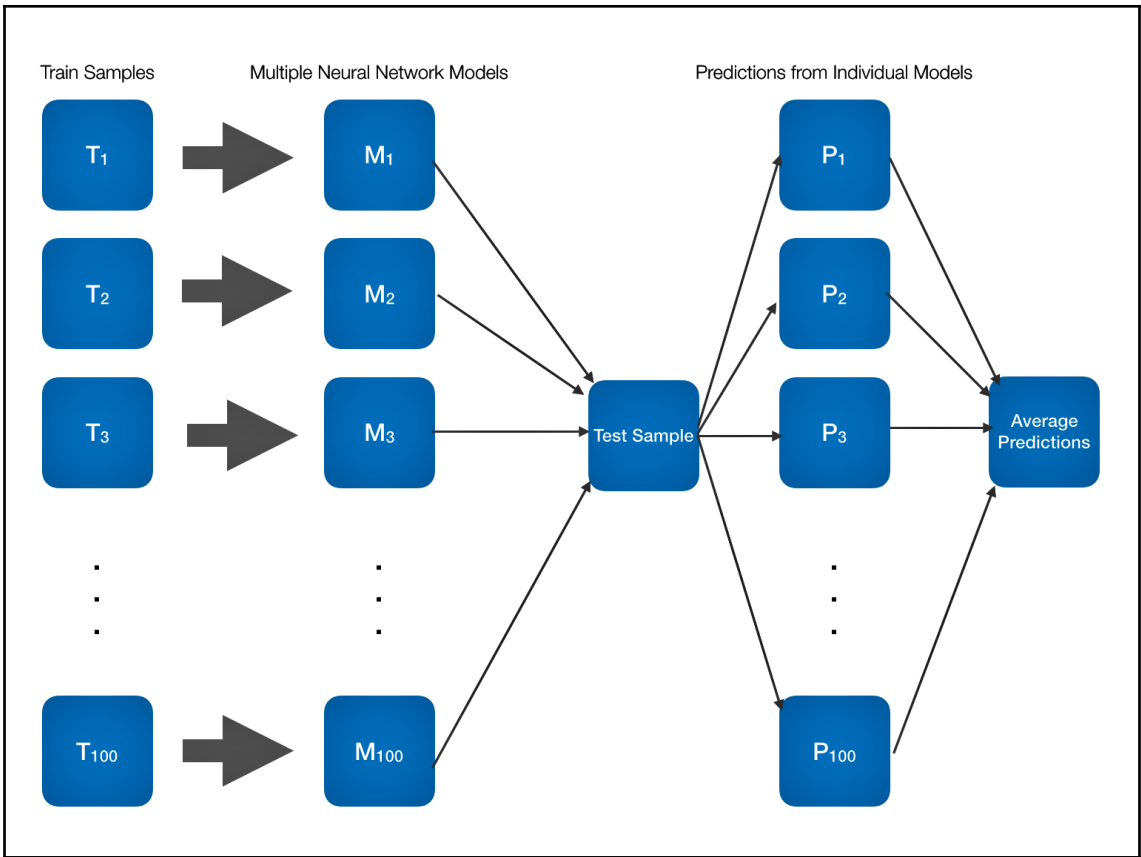
---

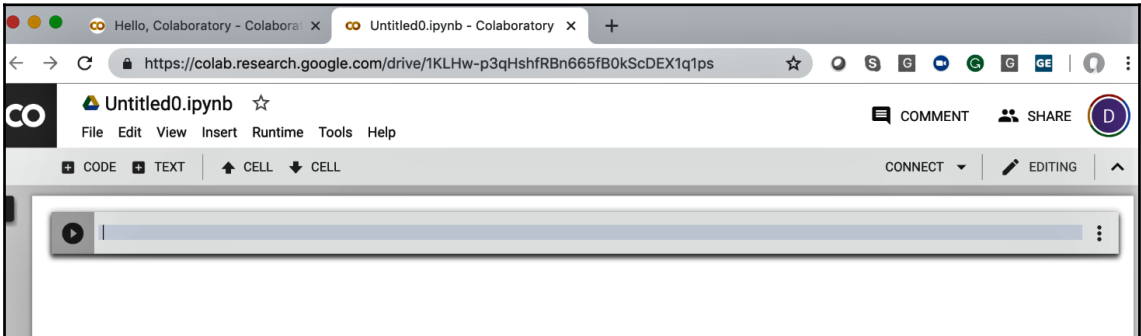
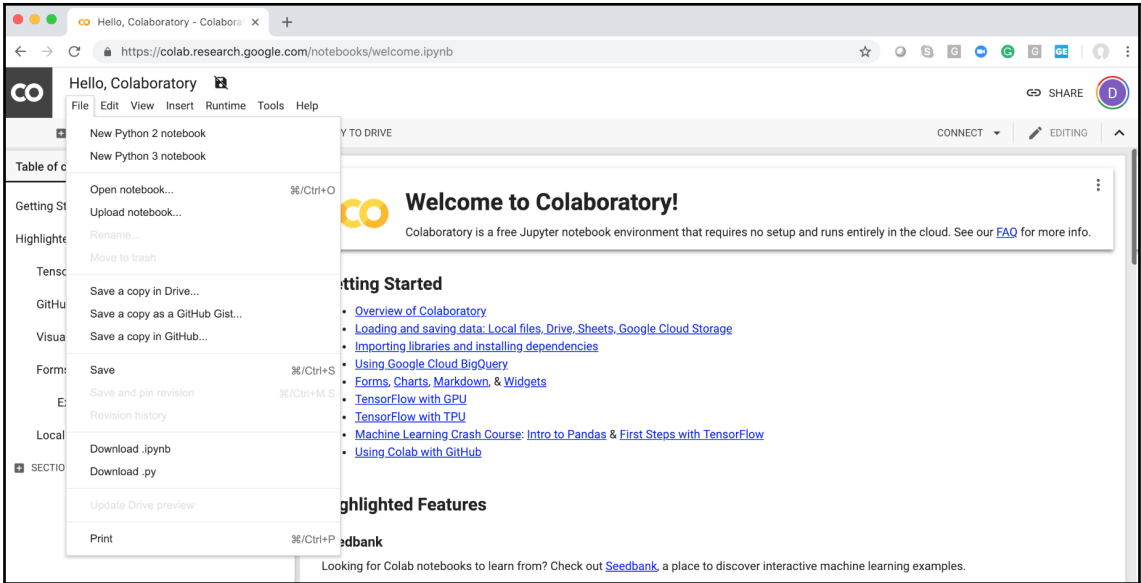
# Chapter 9: Homogeneous Ensembles Using Keras

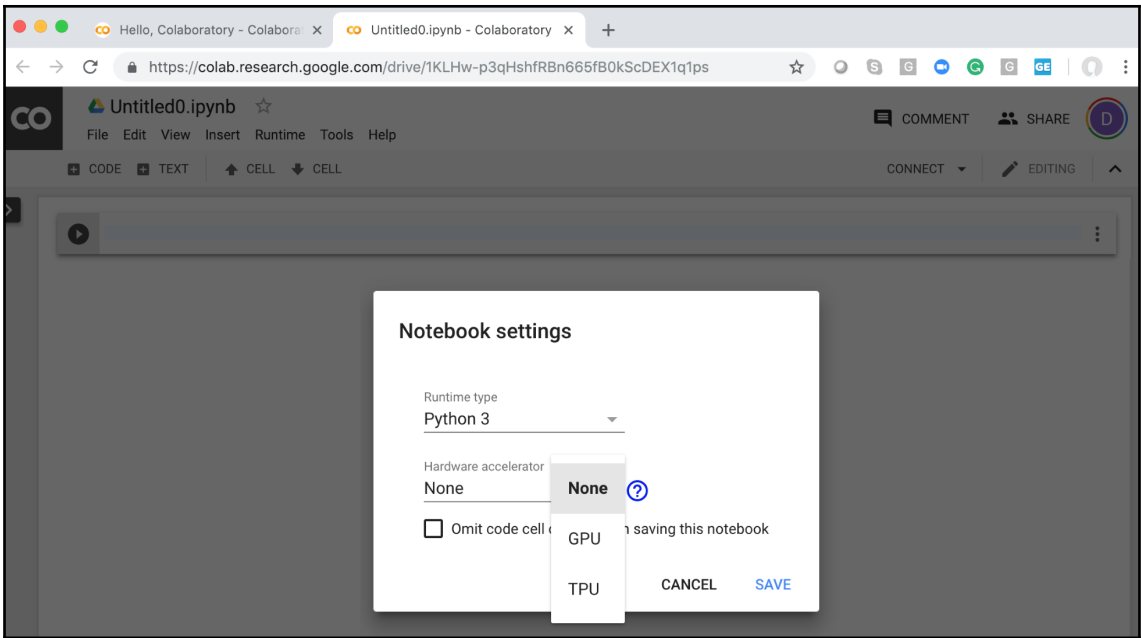


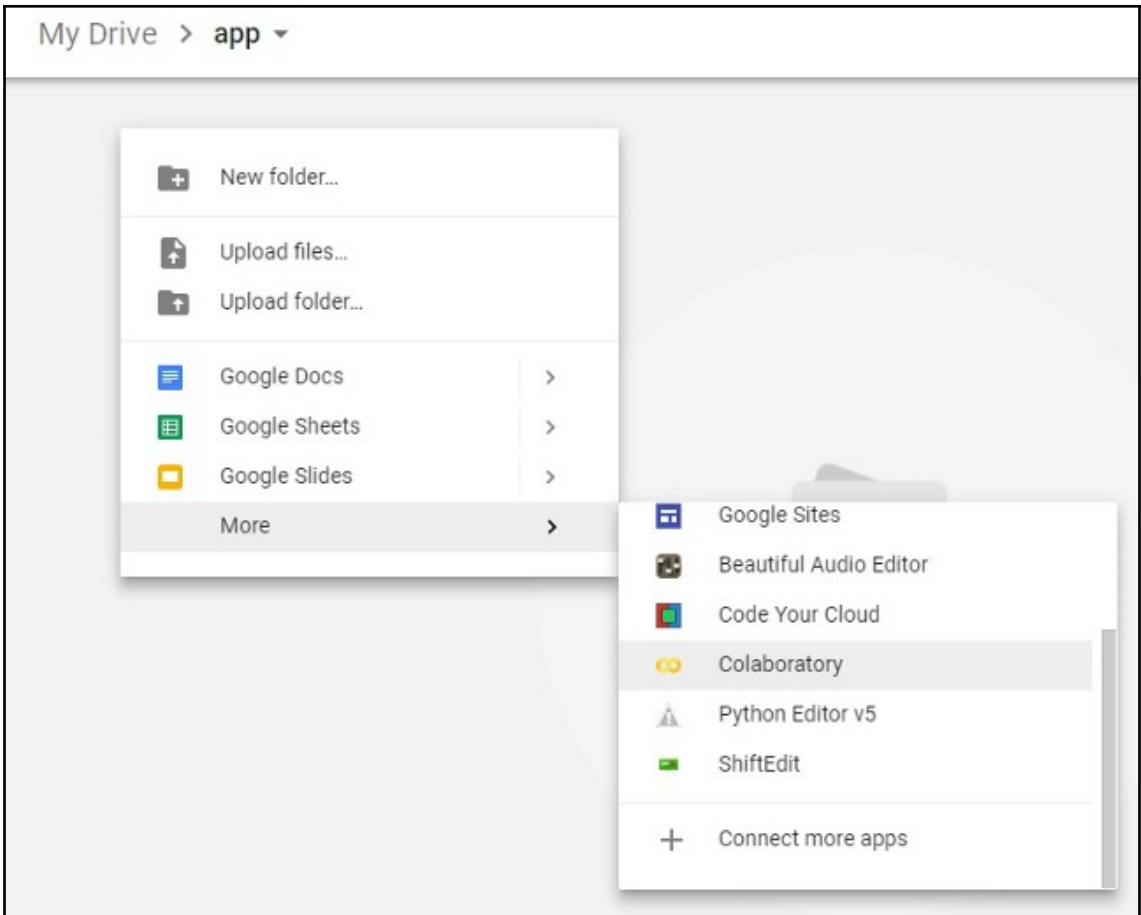
---












... Go to this URL in a browser: [https://accounts.google.com/o/oauth2/auth?client\\_id=947318989803-6bn6qk8qdgf4n4g3pfee6491hc0brc4i.apps.googleusercontent.com&response\\_type=code&scope=https://www.googleapis.com/auth/drive](https://accounts.google.com/o/oauth2/auth?client_id=947318989803-6bn6qk8qdgf4n4g3pfee6491hc0brc4i.apps.googleusercontent.com&response_type=code&scope=https://www.googleapis.com/auth/drive)  
Enter your authorization code:



## Sign in

Please copy this code, switch to your application and paste it there:

4/0ABIoS1CdCSV4JRpt\_Qfburp7vqhCKb0DI8Ce3JXqua   
dEnJjefvhnXQ

```
X_train shape (42000, 1024)
y_train shape (42000, 10)
X_test shape (18000, 1024)
y_test shape (18000, 10)
Train data matrix shape (42000, 1024)
Test data matrix shape (18000, 1024)
```



Before one-hot encoding:

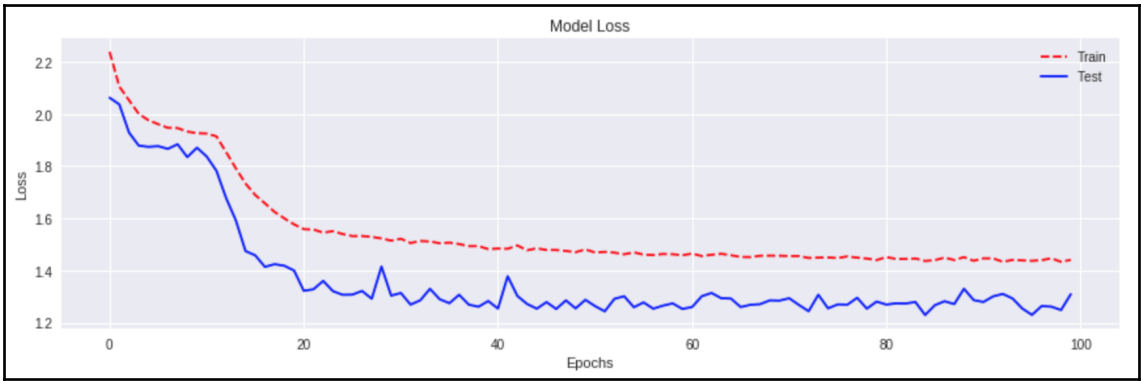
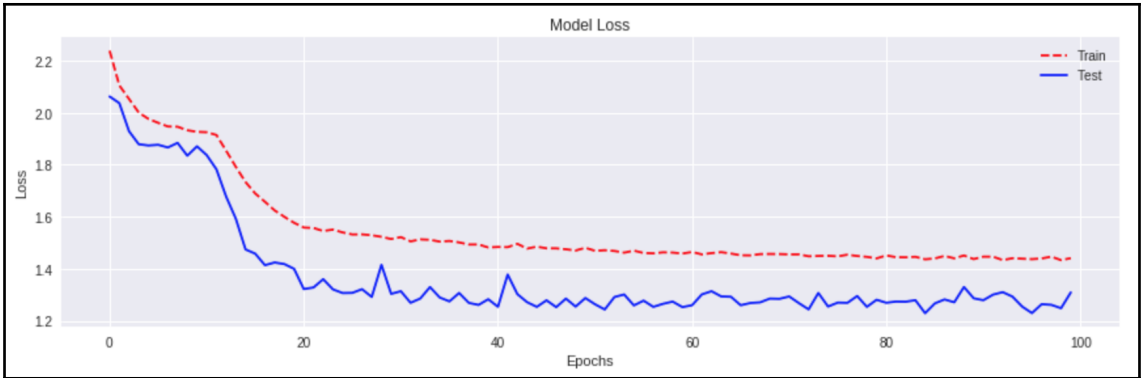
Shape of Y\_TRAIN before one-hot encoding: (42000, 10)

Shape of Y\_TEST before one-hot encoding: (18000, 10)

After one-hot encoding:

Shape of Y\_TRAIN after one-hot encoding: (42000, 10, 10)

Shape of Y\_TEST after one-hot encoding: (18000, 10, 10)

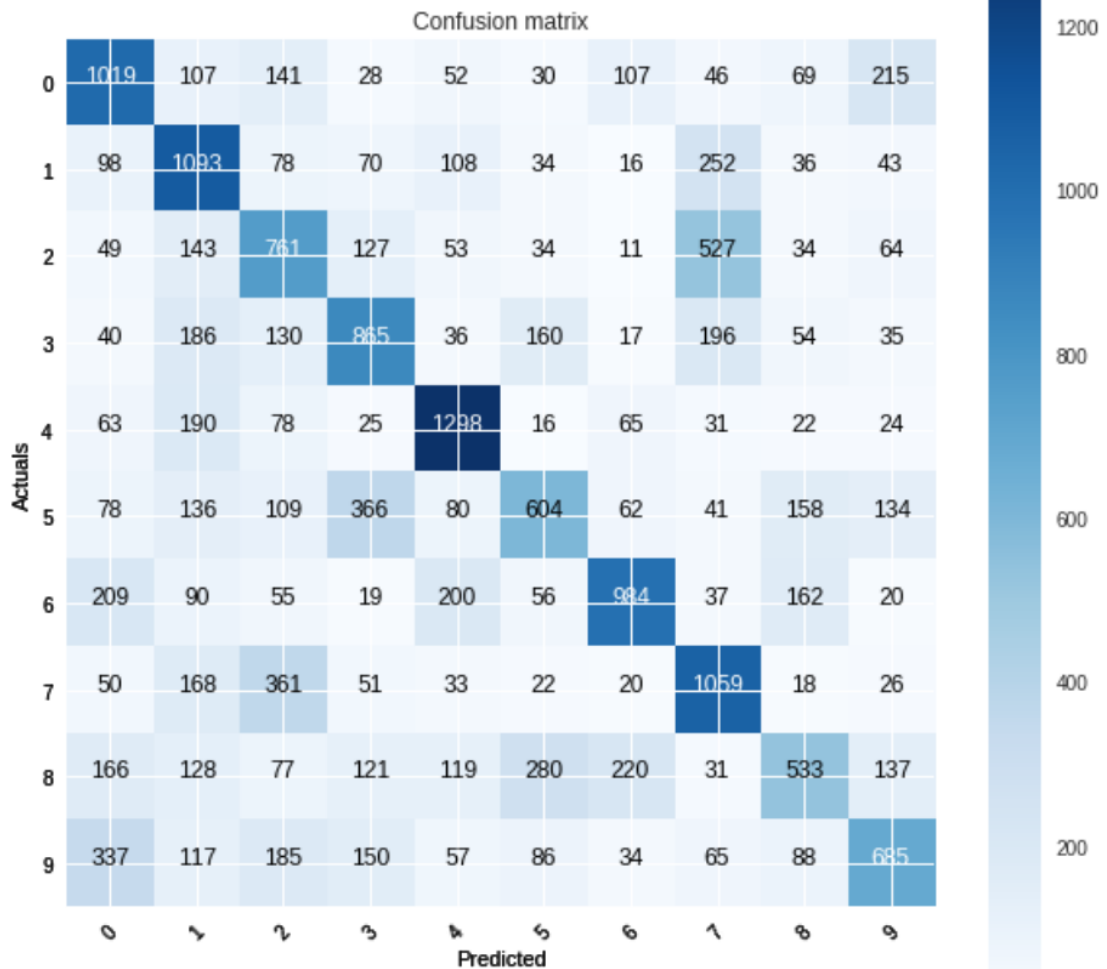




```

[[1019 107 141 28 52 30 107 46 69 215]
 [ 98 1093 78 70 108 34 16 252 36 43]
 [ 49 143 761 127 53 34 11 527 34 64]
 [ 40 186 130 865 36 160 17 196 54 35]
 [ 63 190 78 25 1298 16 65 31 22 24]
 [ 78 136 109 366 80 604 62 41 158 134]
 [ 209 90 55 19 200 56 984 37 162 20]
 [ 50 168 361 51 33 22 20 1059 18 26]
 [ 166 128 77 121 119 280 220 31 533 137]
 [ 337 117 185 150 57 86 34 65 88 685]]

```



---

## Chapter 10: Heterogeneous Ensemble Classifiers Using H2O

```
Connecting to H2O server at http://127.0.0.1:54321... successful.  
H2O cluster uptime:          02 secs  
H2O cluster timezone:       Etc/UTC  
H2O data parsing timezone:  UTC  
H2O cluster version:        3.22.1.2  
H2O cluster version age:    4 hours and 18 minutes  
H2O cluster name:           H2O_from_python_unknownUser_zhwtlv  
H2O cluster total nodes:    1  
H2O cluster free memory:    2.938 Gb  
H2O cluster total cores:    2  
H2O cluster allowed cores:  2  
H2O cluster status:         accepting new members, healthy  
H2O connection url:         http://127.0.0.1:54321  
H2O connection proxy:       None  
H2O internal security:      False  
H2O API Extensions:         XGBoost, Algos, AutoML, Core V3, Core V4  
Python version:             3.6.7 final
```

---

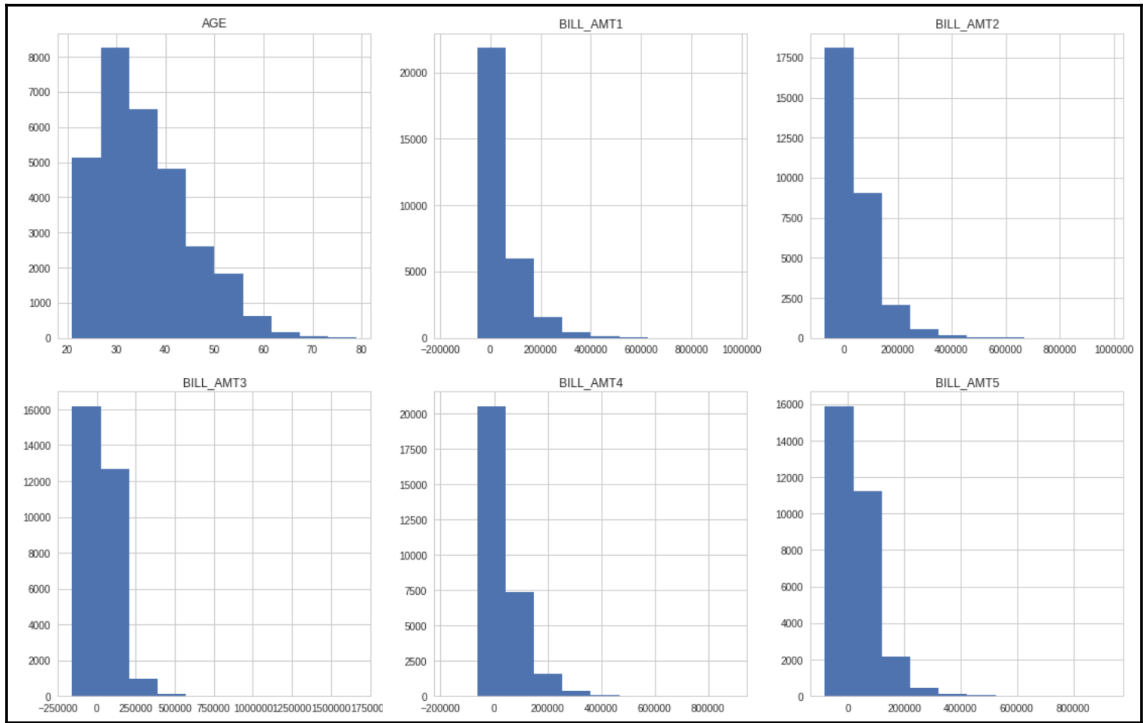
```
{ 'AGE': 'int',  
  'BILL_AMT1': 'int',  
  'BILL_AMT2': 'int',  
  'BILL_AMT3': 'int',  
  'BILL_AMT4': 'int',  
  'BILL_AMT5': 'int',  
  'BILL_AMT6': 'int',  
  'EDUCATION': 'enum',  
  'LIMIT_BAL': 'int',  
  'MARRIAGE': 'enum',  
  'PAY_0': 'enum',  
  'PAY_2': 'enum',  
  'PAY_3': 'enum',  
  'PAY_4': 'enum',  
  'PAY_5': 'enum',  
  'PAY_6': 'enum',  
  'PAY_AMT1': 'int',  
  'PAY_AMT2': 'int',  
  'PAY_AMT3': 'int',  
  'PAY_AMT4': 'int',  
  'PAY_AMT5': 'int',  
  'PAY_AMT6': 'int',  
  'SEX': 'enum',  
  'default.payment.next.month': 'enum' }
```

---

# default.payment.next.month Count

0 23364

1 6636



---

**default.payment.next.month SEX nrow**

0	1	9015
0	2	14349
1	1	2873
1	2	3763

**default.payment.next.month EDUCATION nrow**

0	0	14
0	1	8549
0	2	10700
0	3	3680
0	4	116
0	5	262
0	6	43
1	1	2036
1	2	3330
1	3	1237

**default.payment.next.month MARRIAGE nrow**

0	0	49
0	1	10453
0	2	12623
0	3	239
1	0	5
1	1	3206
1	2	3341
1	3	84

LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6
20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0
120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261
50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547
50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024
100000	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567
140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719
260000	2	1	2	51	-1	-1	-1	-1	-1	2	12261	21670	9966	8517	22287	13668
70000	1	2	2	30	1	2	2	0	0	2	65802	67369	65701	66782	36137	36894
250000	1	1	2	29	0	0	0	0	0	0	70887	67060	63561	59696	56875	55512
50000	2	3	3	23	1	2	0	0	0	0	50614	29173	28116	28771	29531	30211
LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6
90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549
50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131
500000	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944
20000	1	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0	0	13007	13912
200000	2	3	2	34	0	0	2	0	0	-1	11073	9787	5535	2513	1828	3731
630000	2	2	2	41	-1	0	-1	-1	-1	-1	12137	6500	6500	6500	6500	2870
130000	2	3	2	39	0	0	0	0	0	-1	38358	27688	24489	20616	11802	930
60000	2	2	2	22	0	0	0	0	0	-1	15054	9806	11068	6026	-28335	18660
280000	1	1	2	31	-1	-1	2	-1	0	-1	498	9075	4641	9976	17976	9477
100000	2	3	3	43	0	0	0	0	0	0	61559	51163	43824	39619	35762	33258

	alpha	lambda	model_ids	auc
0	[0.1]	[0.001]	GLM_grid_model_42	0.7713781111059813
1	[0.001]	[0.001]	GLM_grid_model_36	0.7709312585464846
2	[0.001]	[0.001]	GLM_grid_model_36	0.7709312585464846
3	[0.001]	[0.001]	GLM_grid_model_29	0.7709312585464846
4	[0.01]	[0.001]	GLM_grid_model_33	0.7708762163954489
5	[0.01]	[0.001]	GLM_grid_model_12	0.7708762163954489
6	[0.05]	[0.001]	GLM_grid_model_23	0.7708584425688997
7	[0.05]	[0.001]	GLM_grid_model_4	0.7708584425688997
8	[0.1]	[0.001]	GLM_grid_model_37	0.7707119599447292
9	[0.1]	[0.001]	GLM_grid_model_21	0.7707119599447292
10	[0.1]	[0.001]	GLM_grid_model_15	0.7707119599447292
11	[0.001]	[0.01]	GLM_grid_model_39	0.7701545977486606
12	[0.001]	[0.01]	GLM_grid_model_40	0.7701448601663852
13	[1.0E-4]	[0.01]	GLM_grid_model_43	0.7701052782799948
14	[0.01]	[0.01]	GLM_grid_model_41	0.7699872479561195
15	[1.0]	[0.001]	GLM_grid_model_20	0.7696323643156521
16	[1.0]	[0.001]	GLM_grid_model_5	0.7696323643156521
17	[0.001]	[0.01]	GLM_grid_model_31	0.7694406029005699
18	[0.001]	[0.01]	GLM_grid_model_38	0.7694406029005699
19	[0.01]	[0.01]	GLM_grid_model_16	0.7693231848005969
20	[0.01]	[0.01]	GLM_grid_model_25	0.7693231848005969
21	[0.05]	[0.01]	GLM_grid_model_2	0.7689357888486935
22	[0.1]	[0.01]	GLM_grid_model_34	0.7685357488620949
23	[0.1]	[0.01]	GLM_grid_model_14	0.7685357488620949
24	[0.001]	[0.1]	GLM_grid_model_17	0.7618706605730926
25	[0.01]	[0.1]	GLM_grid_model_3	0.7614101006426495
26	[0.01]	[0.1]	GLM_grid_model_32	0.7614101006426495
27	[0.05]	[0.1]	GLM_grid_model_22	0.7561812316279662

```

ModelMetricsBinomialGLM: glm
** Reported on cross-validation data. **

MSE: 0.13760225425088768
RMSE: 0.37094777833394244
LogLoss: 0.4394365568279339
Null degrees of freedom: 21100
Residual degrees of freedom: 21039
Null deviance: 22491.611606980696
Residual deviance: 18545.101571252468
AIC: 18669.101571252468
AUC: 0.7713781111059813
pr_auc: 0.5420207583434027
Gini: 0.5427562222119626
Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.2298965437009241:

```

	<b>0</b>	<b>1</b>	<b>Error</b>	<b>Rate</b>
0	13661.0	2697.0	0.1649	(2697.0/16358.0)
1	1995.0	2748.0	0.4206	(1995.0/4743.0)
Total	15656.0	5445.0	0.2224	(4692.0/21101.0)

```

Model Details
=====
H2ORandomForestEstimator : Distributed Random Forest
Model Key: RF_gridsearch_model_29

ModelMetricsBinomial: drf
** Reported on train data. **

MSE: 0.13794252425933623
RMSE: 0.37140614461709737
LogLoss: 0.439212061123055
Mean Per-Class Error: 0.29017439406262435
AUC: 0.7775937549400496
pr_auc: 0.5444830285801324
Gini: 0.5551875098800991
Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.2533025703444073:

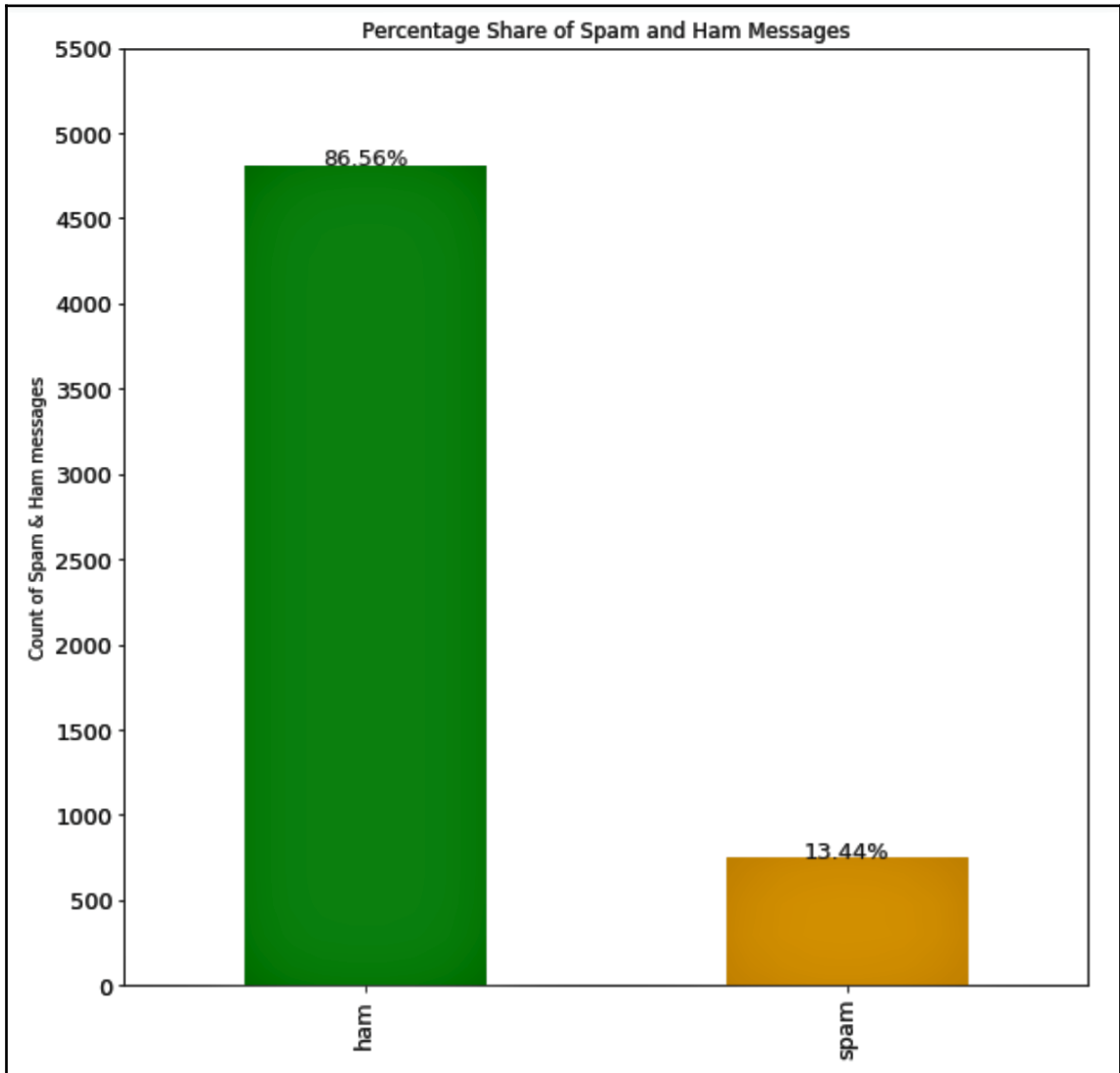
```

	<b>0</b>	<b>1</b>	<b>Error</b>	<b>Rate</b>
0	13744.0	2605.0	0.1593	(2605.0/16349.0)
1	2027.0	2714.0	0.4275	(2027.0/4741.0)
Total	15771.0	5319.0	0.2196	(4632.0/21090.0)



---

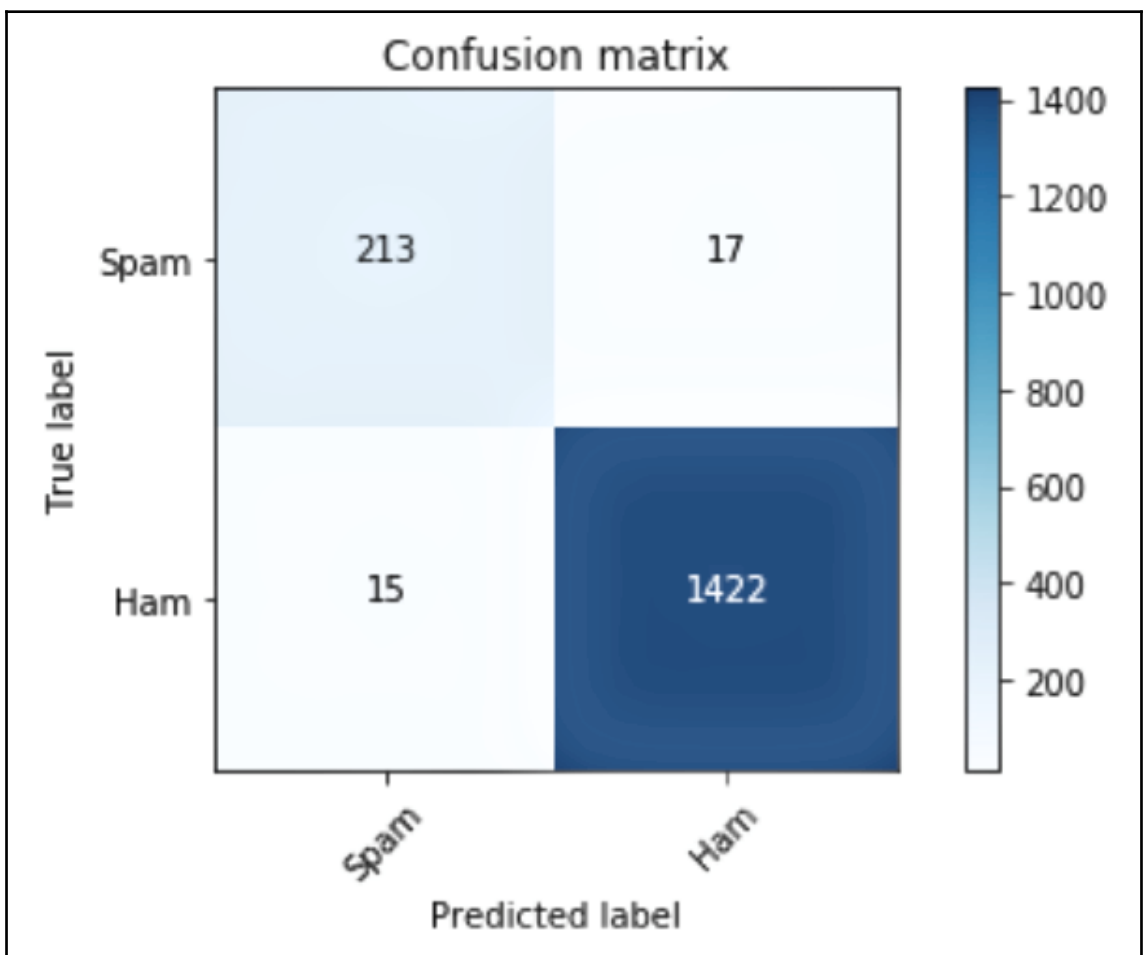
# Chapter 11: Heterogeneous Ensemble for Text Classification Using NLP

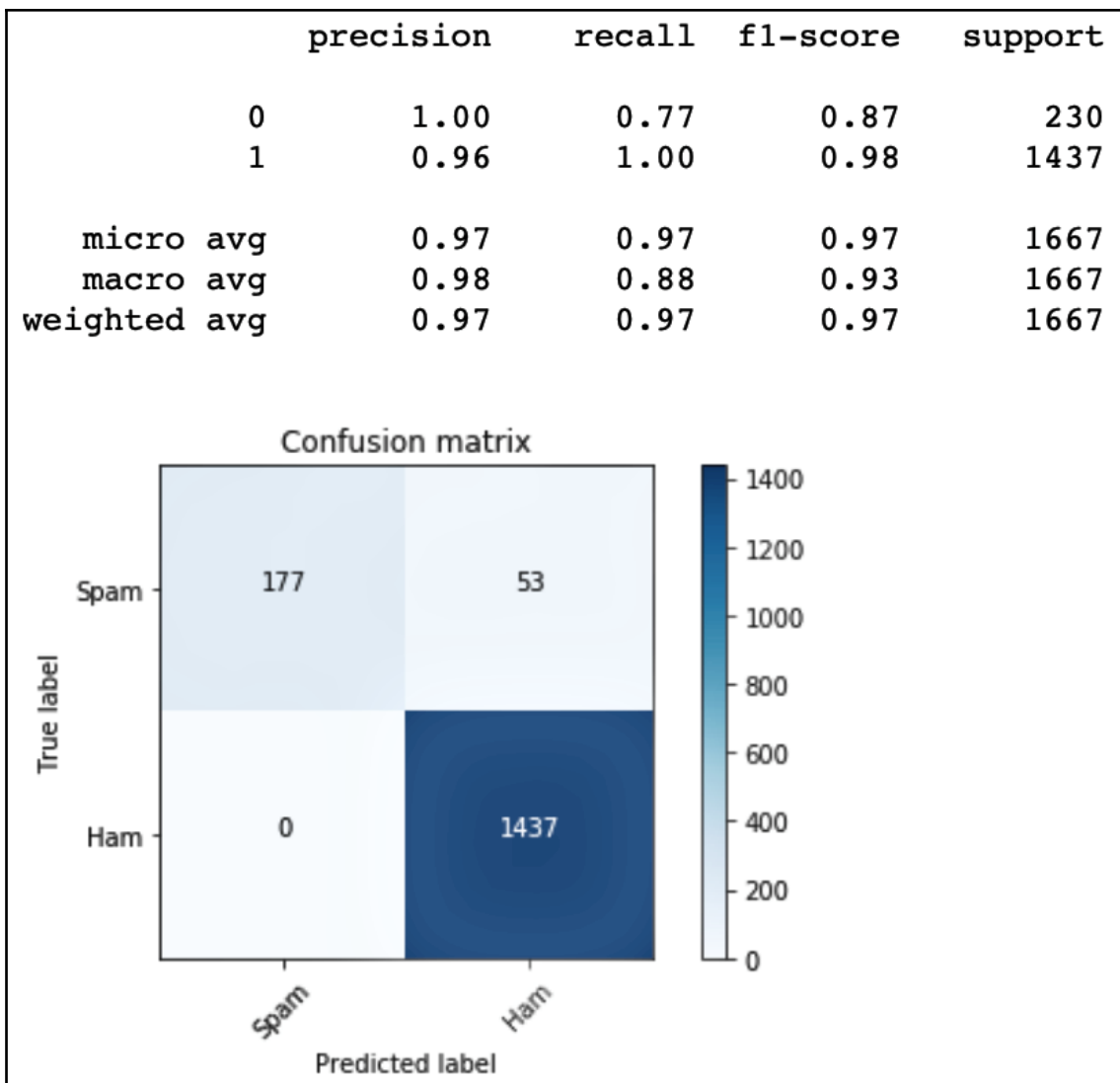


The accuracy for the training data is 0.9940858832604783  
The accuracy for the testing data is 0.9808038392321535

---

	precision	recall	f1-score	support
0	0.93	0.93	0.93	230
1	0.99	0.99	0.99	1437
micro avg	0.98	0.98	0.98	1667
macro avg	0.96	0.96	0.96	1667
weighted avg	0.98	0.98	0.98	1667

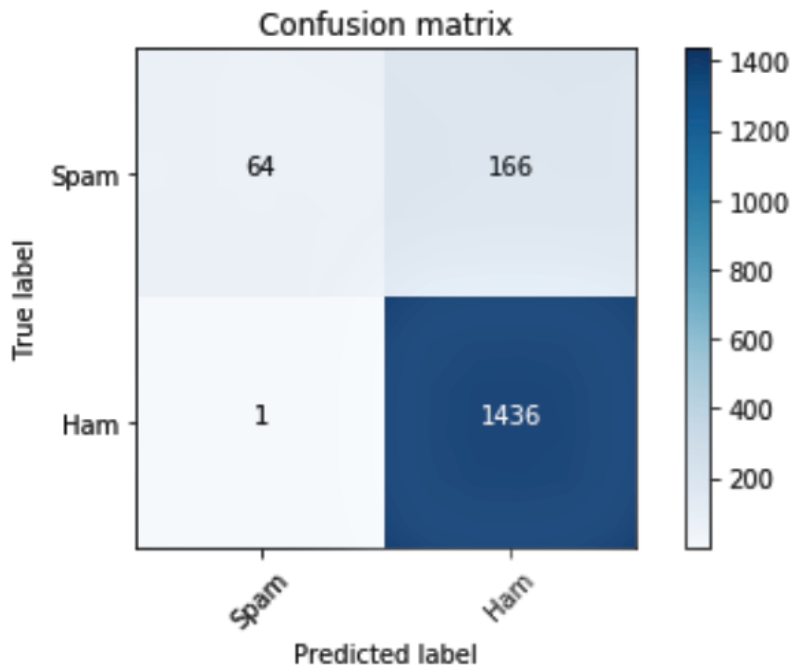




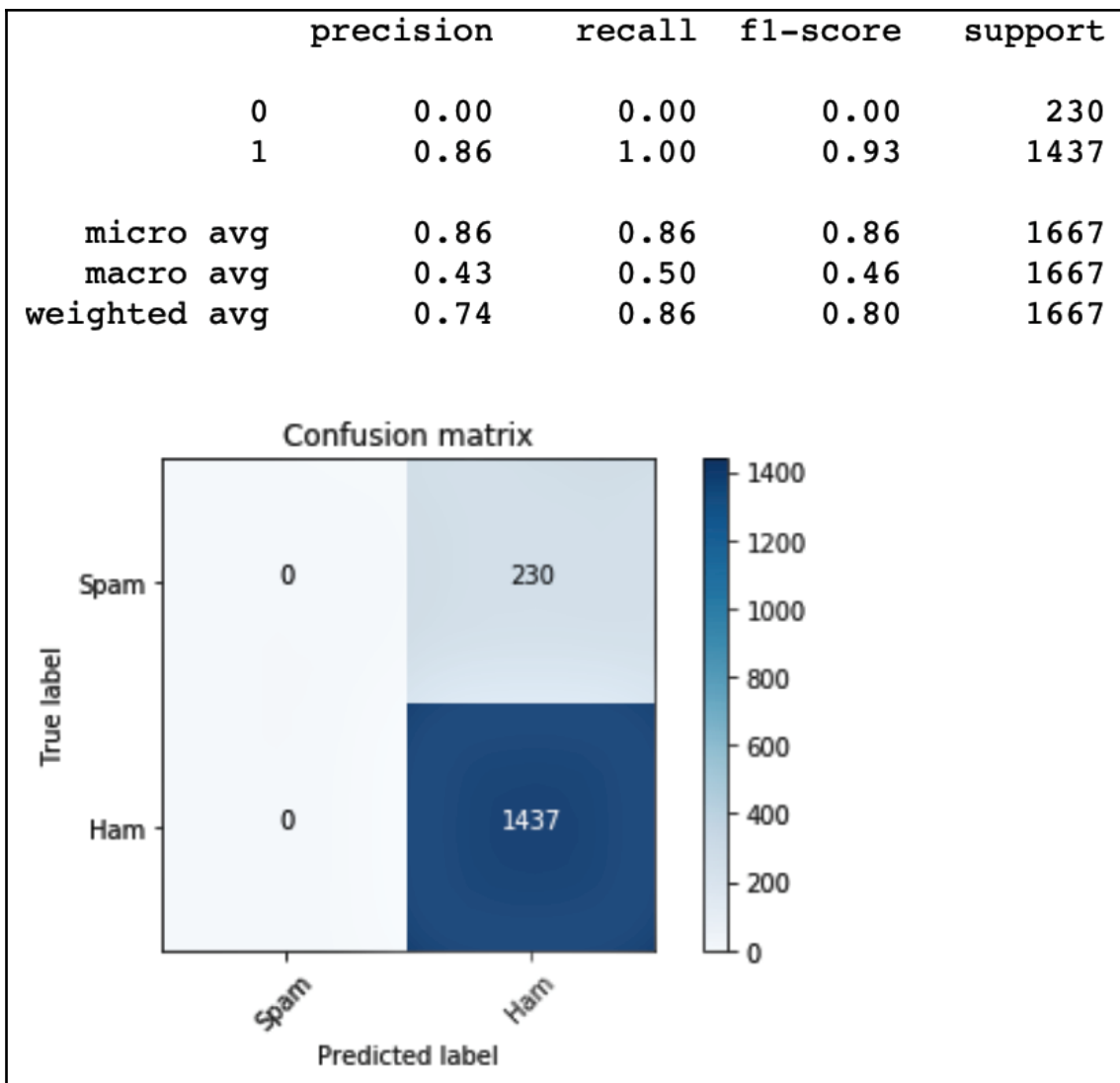
The best parameters {'C': 10}  
The best score 0.8794034456158395

---

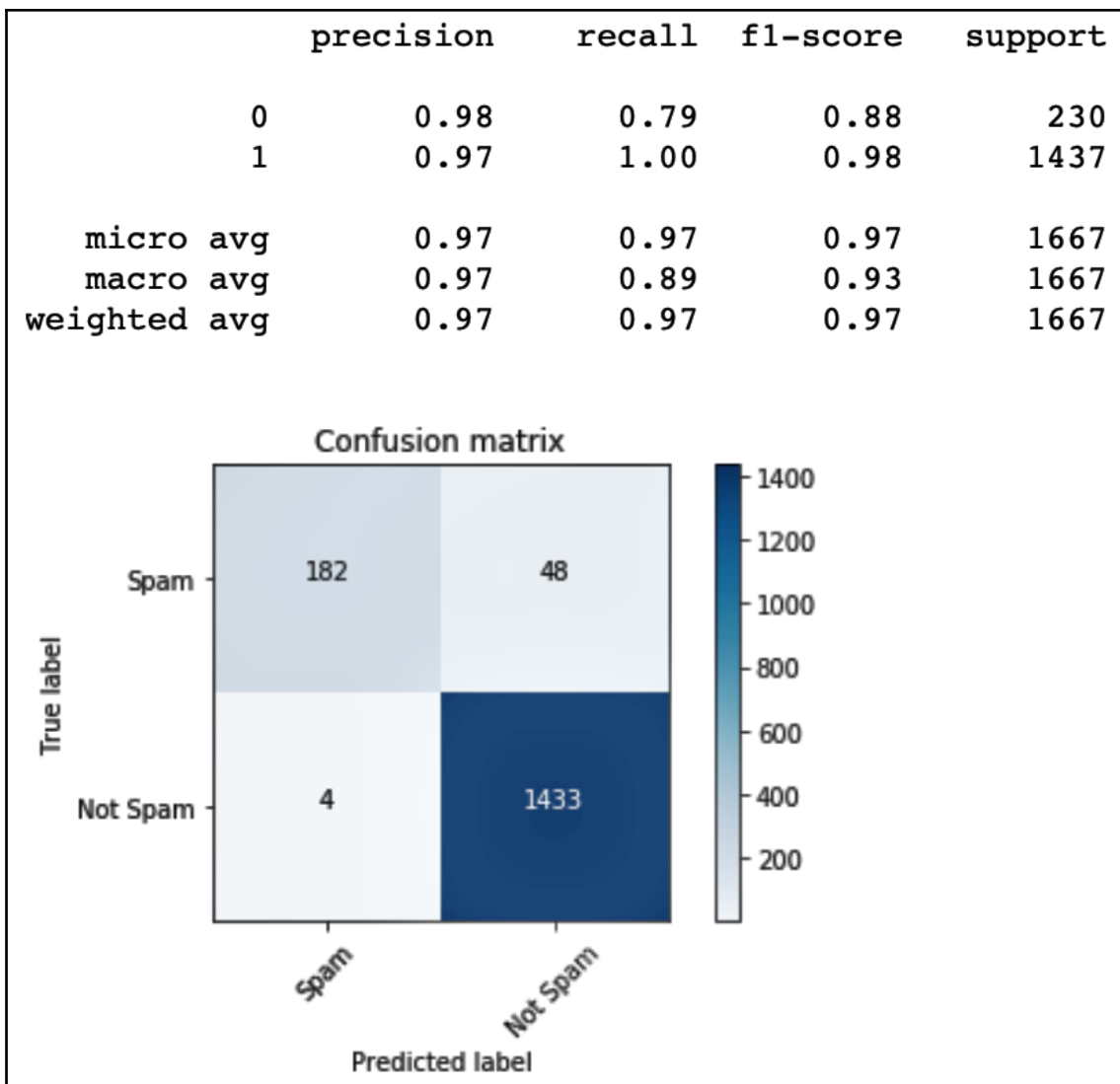
	precision	recall	f1-score	support
0	0.98	0.28	0.43	230
1	0.90	1.00	0.95	1437
micro avg	0.90	0.90	0.90	1667
macro avg	0.94	0.64	0.69	1667
weighted avg	0.91	0.90	0.87	1667

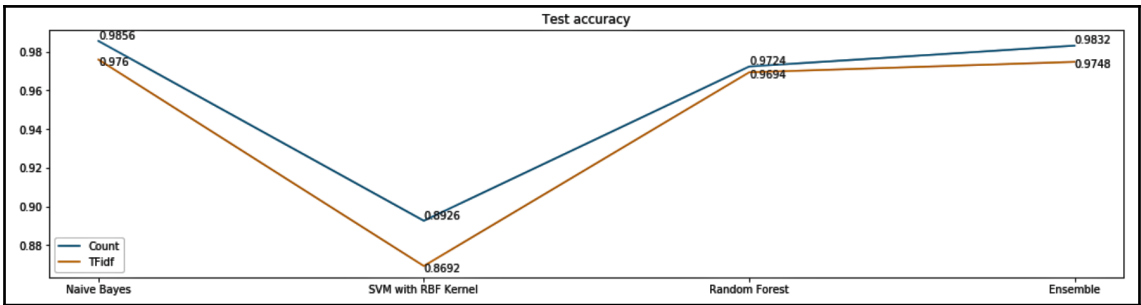
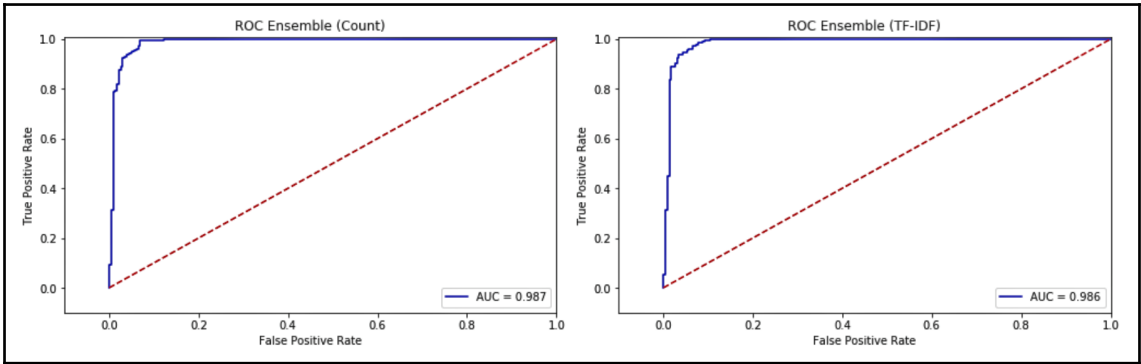
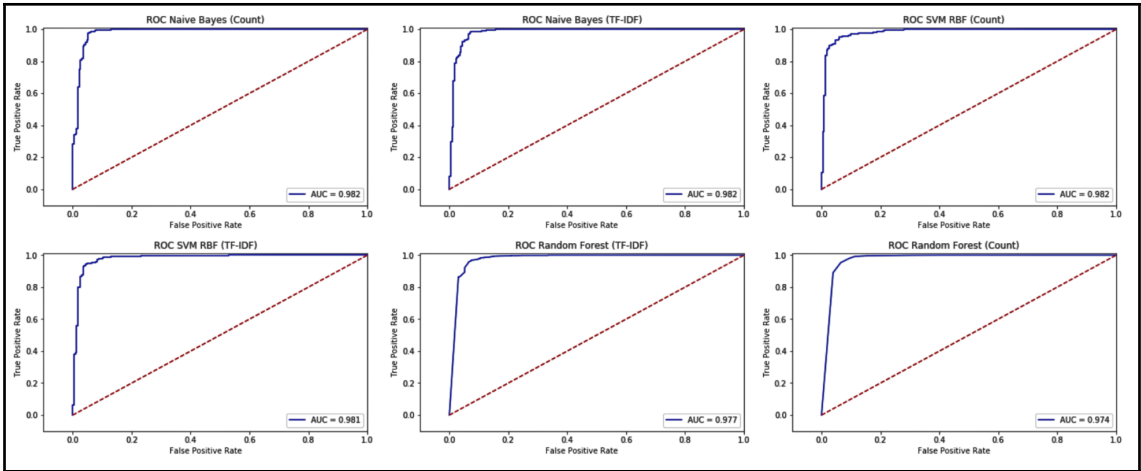


The best parameters {'C': 0.001}  
The best score 0.8670609411159681



The best parameters {'criterion': 'gini', 'max\_depth': None, 'max\_leaf\_nodes': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 3, 'oob\_score': True}  
The best score 0.9748007199794292





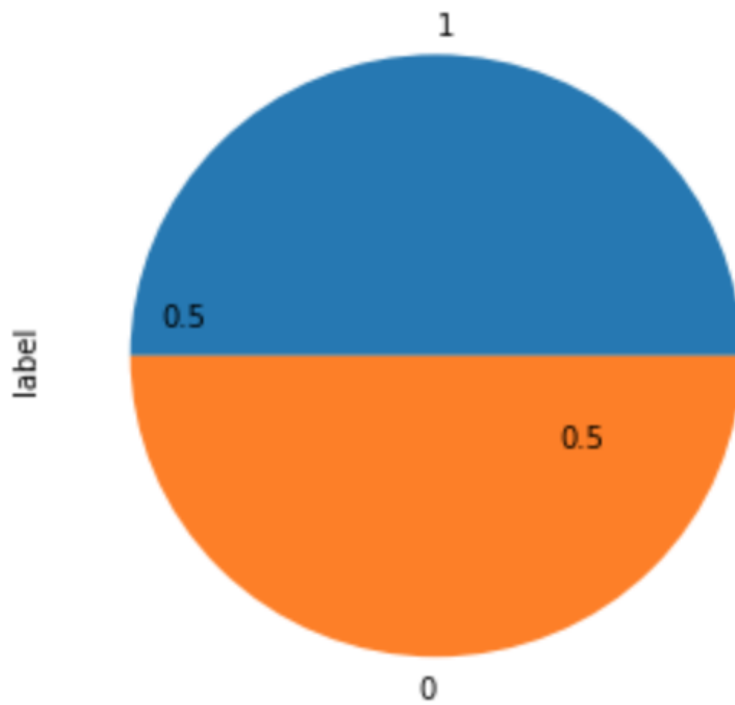
	label	text
0	positive	assume nothing . \nthe phrase is perhaps one o...
1	positive	plot : derek zoolander is a male model . \nhe ...
2	positive	i actually am a fan of the original 1961 or so...
3	positive	a movie that's been as highly built up as the ...
4	positive	" good will hunting " is two movies in one : ...
	label	text
995	negative	synopsis : when a meteorite crashlands in the ...
996	negative	it's now the anniversary of the slayings of ju...
997	negative	coinciding with the emerging popularity of mov...
998	negative	and now the high-flying hong kong style of fil...
999	negative	battlefield long , boring and just plain stupi...

	label	text
<b>879</b>	negative	capsule : the running gag pair of characters f...
<b>310</b>	positive	i rented this movie with very high hopes . \nt...
<b>228</b>	negative	the scene at the end of 1989's `dead poets soc...
<b>891</b>	negative	the makers of spawn have created something alm...
<b>753</b>	positive	robert benton has assembled a stellar , mature...
<b>689</b>	negative	tom dicillo directs this superficial comedy ab...
<b>106</b>	negative	another f <code>txt_sentoken.zip</code> quirky comedy from...
<b>130</b>	negative	five years after his directorial debut based o...
<b>382</b>	positive	another 'independent film' , this comedy , whi...
<b>222</b>	positive	after the average mouse hunt , the silly small...

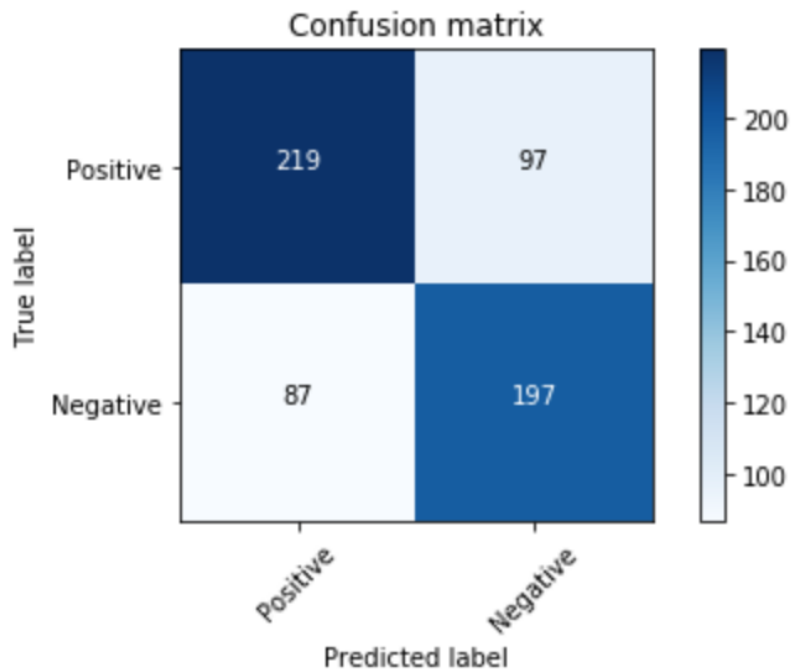


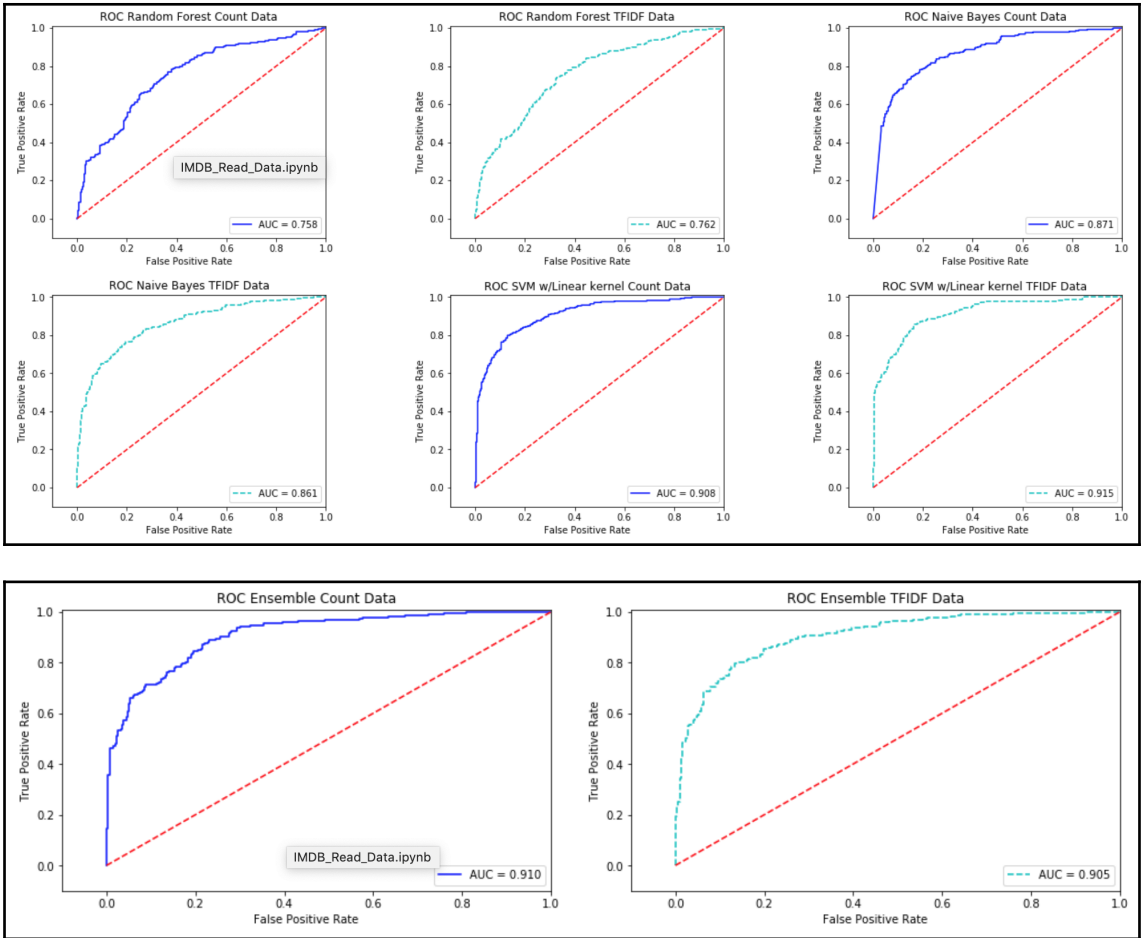
---

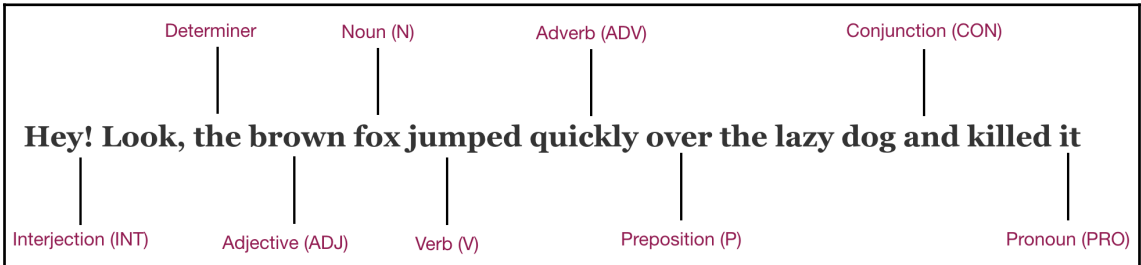
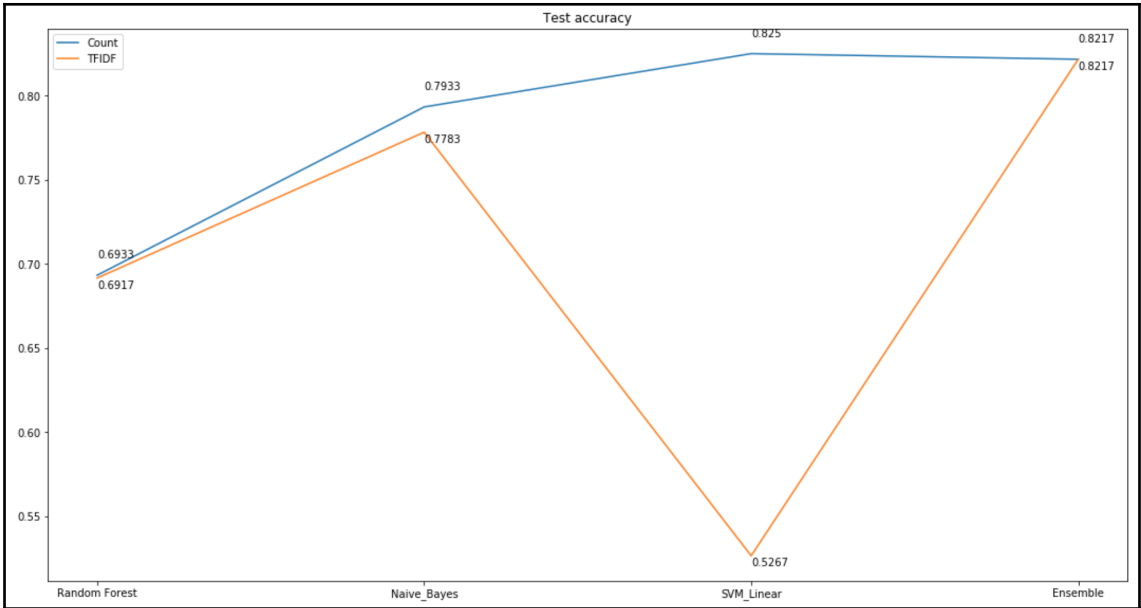
% Share of the Positive and Negative reviews in the dataset



	precision	recall	f1-score	support
0	0.72	0.69	0.70	316
1	0.67	0.69	0.68	284
micro avg	0.69	0.69	0.69	600
macro avg	0.69	0.69	0.69	600
weighted avg	0.69	0.69	0.69	600







---

```
[ 'star',  
  'war',  
  'came',  
  'twenty',  
  'year',  
  'ago',  
  'image',  
  'traveling',  
  'throughout',  
  'star' ]
```

---

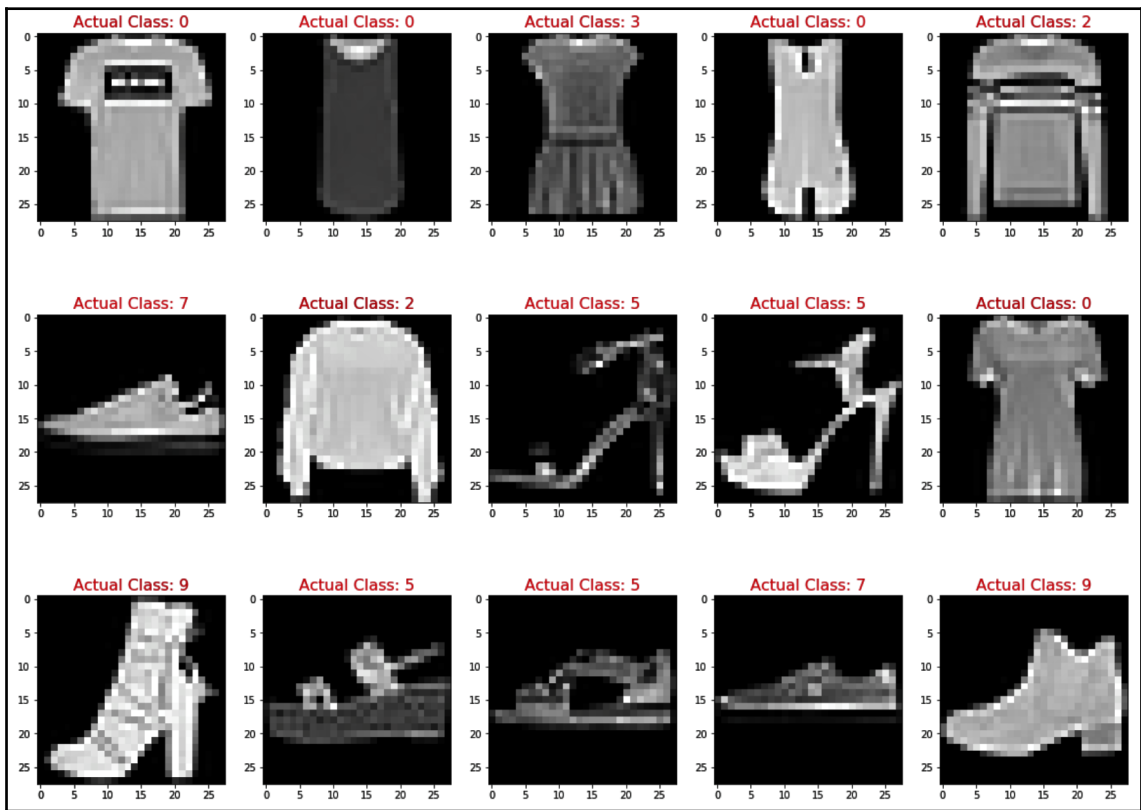
```
[('star', 'JJ'),
 ('war', 'NN'),
 ('came', 'VBD'),
 ('twenty', 'CD'),
 ('year', 'NN'),
 ('ago', 'RB'),
 ('image', 'NN'),
 ('traveling', 'VBG'),
 ('throughout', 'IN'),
 ('star', 'NN')]
```

---

# Chapter 12: Homogenous Ensemble for Multiclass Classification Using Keras

```
x_train shape: (60000, 28, 28) /n y_train shape: (60000,)
```

```
array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9], dtype=uint8)
```



```

Epoch 1/10
60000/60000 [=====] - 21s 351us/step - loss: 0.6613 - acc: 0.7526
Epoch 2/10
60000/60000 [=====] - 13s 214us/step - loss: 0.4929 - acc: 0.8265
Epoch 3/10
60000/60000 [=====] - 13s 213us/step - loss: 0.4746 - acc: 0.8398
Epoch 4/10
60000/60000 [=====] - 13s 214us/step - loss: 0.4096 - acc: 0.8574
Epoch 5/10
60000/60000 [=====] - 13s 213us/step - loss: 0.3872 - acc: 0.8656
Epoch 6/10
60000/60000 [=====] - 13s 214us/step - loss: 0.3854 - acc: 0.8675
Epoch 7/10
60000/60000 [=====] - 13s 213us/step - loss: 0.3710 - acc: 0.8718
Epoch 8/10
60000/60000 [=====] - 13s 213us/step - loss: 0.3838 - acc: 0.8680
Epoch 9/10
60000/60000 [=====] - 13s 213us/step - loss: 0.3645 - acc: 0.8743
Epoch 10/10
60000/60000 [=====] - 13s 213us/step - loss: 0.3596 - acc: 0.8758
10000/10000 [=====] - 4s 431us/step
[0.38200154910087586, 0.8627]
Epoch 1/10
60000/60000 [=====] - 22s 359us/step - loss: 0.6461 - acc: 0.7645
Epoch 2/10
60000/60000 [=====] - 13s 214us/step - loss: 0.4749 - acc: 0.8359
Epoch 3/10
60000/60000 [=====] - 13s 214us/step - loss: 0.4325 - acc: 0.8506
Epoch 4/10
60000/60000 [=====] - 13s 215us/step - loss: 0.4129 - acc: 0.8590

```



---

	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>
<b>0</b>	0.8683	0.869687	0.8683
<b>1</b>	0.8435	0.849763	0.8435
<b>2</b>	0.8573	0.859797	0.8573
<b>3</b>	0.8654	0.867831	0.8654
<b>4</b>	0.8687	0.869421	0.8687
<b>5</b>	0.8453	0.859606	0.8453
<b>6</b>	0.8656	0.866652	0.8656



	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>Models</b>
<b>0</b>	0.851	0.856075	0.851	Model 1
<b>1</b>	0.8601	0.861578	0.8601	Model 2
<b>2</b>	0.8531	0.855948	0.8531	Model 3
<b>3</b>	0.8331	0.843468	0.8331	Model 4
<b>4</b>	0.8467	0.846121	0.8467	Model 5
<b>5</b>	0.8523	0.85146	0.8523	Model 6
<b>6</b>	0.8353	0.837242	0.8353	Model 7
<b>7</b>	0.8675	0.8675	0.8675	Ensemble Model

