DATA7202 : Assessment 1

Name : Peng Yu
Student ID : 46635884

Q1:

We first start with linear regression. By fitting the linear regression, we arrive at the following result

	coef	std err	t	P>|t|	[0.025	0.975]
radio	0.8734	0.022	39.950	0.000	0.831	0.916
tv	4.1962	0.038	109.153	0.000	4.121	4.272
internet	6.1841	0.040	155.965	0.000	6.106	6.262

First, this is not clear if radio advertisement contributes to sales. In particular, note that the confidence interval is (0.831, 0.916). In other words, there is no statistically significant evidence that the radio coefficient is not zero.
Second, in addition, the tv and the internet seem to contribute directly to sales. The corresponding confidence intervals suggest that these coefficients are positive (and the result is statistically significant).
Finally, regression coefficients suggest that the most beneficial domain for advertisement is the internet, the second on is tv, and the last one is radio.

```
regression mean squared error:  0.10828680248608381
random forest loss:  0.10863824674241702
```

The MSE of Linear Regression is 0.108286
The MSE of Random Forest is 0.108638

Q2

In my opinion, this is not a good model. Getting good predictors on a subset can only mean that there is a good prediction on this set, but not on the whole data set. Therefore, the prediction error obtained by this model is only fitting for this subset, not necessarily for the whole data set. Moreover, if the prediction is too good, and the prediction model is used to predict the whole data set, it is possible to form an overfitting phenomenon. To sum up, I cannot expect to obtain the true prediction error.

Q3:

θ is an unknown vector of parameters that take values in the parameter space Θ.
F is a normal distribution with mean μ and standard deviation of $\boldsymbol{\sigma}$**.**
Here $\theta = \{\mu, \boldsymbol{\sigma}\}$ and $\Theta = \{\mathbb{R}, (0, +\infty)\}$.

Q4:

$$\mathbb{E}_{\mathcal{T}} Loss_{\mathcal{T}}(g) \triangleq \mathbb{E}_{\mathcal{T}}\left[\frac{\sum_{i=1}^{m} 1\{g(x_i) \neq y_i\}}{m}\right]$$

∵Training set is a sample from distribution $\mathcal{D}$

$$= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\mathcal{D}} 1\{g(x) \neq g^*(x)\}$$

$$= \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}_{x\sim\mathcal{D}}(g(x) \neq g^*(x))$$

$$\triangleq \frac{1}{m}\sum_{i=1}^{m} Loss_{\mathcal{D}}(g)$$

$$= \frac{1}{m} * m * Loss_{\mathcal{D}}(g)$$

$$= Loss_{\mathcal{D}}(g)$$

Q5:

```
beta_0:  0.0
beta_1:  0.6
********************
beta_0:  1.8
beta_1:  3.510833468576701e-17
********************
Model1 mean squared error loss:  1.64
Model1 mean squared absolute error loss:  1.1600000000000001
Model1 mean squared L1.5 loss:  1.36348016266711
********************
Model2 mean squared error loss:  0.5600000000000002
Model2 mean squared absolute error loss:  0.6400000000000001
Model2 mean squared L1.5 loss:  0.5849006163624495
```

(a)

Model	β_0	β_1

$Model_1$	0	0.6
$Model_2$	1.8	3.510833*10^-16

(b)

Model	squared error loss	absolute error loss	$L_{1.5} loss$
$Model_1$	1.64	1.16	1.3634480
$Model_2$	0.56	0.64	0.584900

(c)

No matter through squared error calculation, absolute error calculation, or L1.5 calculation, the final error loss of Model 2 is lower than that of Model 1. To sum up, Model 2 is better than Model 1

Q6

(a)

```
AtBat          int64
Hits           int64
HmRun          int64
Runs           int64
RBI            int64
Walks          int64
Years          int64
CAtBat         int64
CHits          int64
CHmRun         int64
CRuns          int64
CRBI           int64
CWalks         int64
League         int64
Division       int64
PutOuts        int64
Assists        int64
Errors         int64
Salary       float64
NewLeague      int64
dtype: object
```

(b)

OneHotEncoder

Advantages: OneHotEncoder solves the problem that classifiers are not good at processing attribute data, and also plays a role in extending features to some extent. Its values are only 0 and 1, and the different types are stored in vertical space.

Disadvantages: When the number of categories is large, the feature space can become very large.

Label Encoding can be useful in some situations, but the situations are very restrictive. For example, if we have [dog,cat,dog,mouse,cat], we convert it to [1,2,1,3,2]. Here a curious phenomenon arises: the average value of a dog and mouse is a cat.

However, in this data, all the classified data have only dichotomies, so here is better to use LabelEncoder.

(c)

```
In [8]: runcell(0, '/Users/[redacted]/Desktop/computer_science/DATA7202/
assessment_1/Q6.py')
linear regression cross validation error =  116599.01367380246
```

10-Fold Cross-Validation mean squared error = 116599.01367

Q7

$$\int_0^1 \frac{1}{x^2+2x+3}\,dx = \frac{\sqrt{2}\left(-\tan^{-1}\left(\frac{\sqrt{2}}{2}\right)+\tan^{-1}(\sqrt{2})\right)}{2} \approx 0.240300983$$

```
mean =  0.2404 CI = ( 0.23202440283834042 ,  0.2487755971616596 )
```