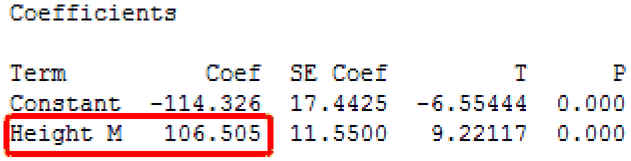
ANSWERS

ANSWER 1

Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant. This statistical control that regression provides is important because it isolates the role of one variable from all of the others in the model. The key to understanding the coefficients is to think of them as slopes, and they're often called slope coefficients.



```
Coefficients

Term          Coef  SE Coef        T      P
Constant  -114.326  17.4425 -6.55444  0.000
Height M   106.505  11.5500  9.22117  0.000
```

In the above example, coefficient of "Height M" variable in a linear regression model is 106.505. It means for every 1 unit increment/decrement in "Height M", the value of target variable will be increased/decreased (since positive) respectively.

Hence, if the linear regression coefficient of a specific variable is 0.54, it means for every additional unit of the predictor the value of the dependent variable will increase by 0.54 unit.

ANSWER 2

The classical data imbalance problem is recognized as one of the major problems in the field of data mining and machine learning as most machine learning algorithms assume that data is equally distributed.

Usually these kinds of problems falls in the category of classification problem. The Class Imbalance Problem is a common problem affecting machine learning due to having disproportionate number of class instances in practice.

To compare solutions, we will use alternative metrics (True Positive, True Negative, False Positive, and False Negative) instead of general accuracy of counting number of mistakes. Due to its prevalence, there are many approaches out there to deal with this problem. They can be generally classified into two major categories of

1) Sampling based

2) Cost function based.

Sampling based can be broken into three major categories:

a) Over sampling

b) Under sampling

c) Hybrid of oversampling and under sampling.

Let,

target(y) ∈ {0, 1}.

0->-ve(100 data-points)

1->+ve(900 data-points)

When you have severely imbalanced(i.e. 0->100pts, 1->900pts) data, then typically the majority class dominate. So, in that situation You can use undersampling/upsampling or oversampling/downsampling.

Undersampling --> Create new dataset with all the 0 class data and among the all 1 class data, randomly sample 100pts and train your model on this dataset. But there is a problem with undersampling as you are creating new dataset which is extremely less than the previous data. So, your model might not work well because you have thrown away many data which contains much information. You should avoid this situation.

Oversampling --> Create new dataset with all the 1 class data and repeat your minority class 9-times, as you have 900pts from +ve class and 100pts from -ve class. This is very simple idea of repeating minority class points. Many many techniques are there in literature you can use. In Logistic Regression, svm. decision tree etc the parameter class_weight= "balanced" works like oversampling.

EX:-

Your original data(100 -ve pts and 900 +ve pts) i.e. the ratio is 1:9 and you split it into 70%(700pts) Train and 30%(100pts) Test randomly then you will have your Train data containing 630 +ve and 70 -ve pts and test data containing 270 +ve and 30 -ve pts. Let, your model classify every pts as -ve on your test data then your accuracy (performance measure/ matrix) will be 90% (270/300). Here, when you have imbalanced data you should go for f1-score, auc etc. instead of accuracy as a performance measure.

ANSWER 3

We should not consider only accuracy as a performance measure as it evaluate only true positive , true Negative and sum total of a model. We have a many performance measures like recall, precision and f1-score. Now, coming to this question statement the classification model with 90% accuracy having high false positive rate. First of all False positive rate is a parameter of error metric derived from confusion matrix. Confusion matrix depends on distinct respective model. Thus, each classification model will have different confusion matrix which turns out to have different False positive rate may be low or high as compared to previous model. Thus, here we can go for various classification model available like as logistics regression, Decision Tree, Neural networks, Random Forest, etc and check false positive rate using confusion matrix for each of the models. On comparison we can conclude which machine learning model or statistical model is best fit having high accuracy and lowest possible false positive rate. A Machine learning paradigm known as ensemble learning can also be used in this condition. Ensemble learning is nothing but the group of different types of machine learning models developed using same training dataset (some feature may or may not differ in the dataset). Ensemble learning is implemented in a technique known as bagging or Bootstrap Aggregating in which several models are trained on a dataset and mean of the output is taken for the test dataset output by each model. Random forest is one such ensemble learning technique which aggregates output of several decision trees to get most appropriate result.

We can also perform predictive analysis on the data such as:

1) Missing value analysis and imputation

2) Feature engineering

3) Normalisation and standardisation

4) Cross validation measures

5) Fine-tuning hyper parameters boosting Algorithms

6) Ensambling

7) Collecting more data

8) Synthesizing more data

ANSWER 4

Multi collinearity is a condition when two or more variables carry almost the same information. This condition will allow the model to be biased towards a variable. On the other hand Naive Bayes algorithm uses the Bayes theorem of probability. It assumes that the presence of one feature does not affect the presence or absence of other feature no matter up to which extent the features are interrelated. So, multi collinearity does NOT affect the Naive Bayes.

Naive Bayes performs well when we have multiple classes and working with text classification. Advantage of Naive Bayes algorithms are:

It is simple and if the conditional independence assumption actually holds, a Naive Bayes classifier will converge quicker than discriminating models like logistic regression, so we would need less training data, and even if the Naive Bayes assumption doesn't hold, It requires less model training time

ANSWER 5

In general, the more trees you use the better get the results. However, the improvement decreases as the number of trees increases, i.e. at a certain point the benefit in prediction performance from learning more trees will be lower than the cost in computation time for learning these additional trees.

Random forests are ensemble methods, and you average over many trees. Similarly, if you want to estimate an average of a real-valued random variable (e.g. the average heigth of a citizen in your country) you can take a sample. The expected variance will decrease as the square root of the sample size, and at a certain point the cost of collecting a larger sample will be higher than the benefit in accuracy obtained from such larger sample.

Typical values for the number of trees is 10, 30 or 100. I think in only very few practical cases more than 300 trees outweights the cost of learning them (well, except maybe if you have a really huge dataset). But if the value for number of trees is not specified, the random forest model takes the default number of trees value as 500 in R and 10 in Python's scikit library.