1)The target class : Absent_Frqcy_Class / I think 'target class' means literally 'the class that is the target of classification. We may analyze data and patterns on certain purpose, and the purpose would be certain subjects which is classified by ML.

(I'll paint the same color with things that have direct correlations.)
Features : 1.Month of absence, 2.Day of the week, 3.Season (The hotter and the farther away the day of the week is from the weekend, the more likely to be absent)
4.Transportation expense, 5.Distance from Residence to Work (The farther away from the workplace, the more likely to be absent(usually transportation cost gets more expensive as the distance to the destination increases))
6.Service time, 7.Age (The longer they worked at their workplace, the lesser their patience (of course this is just my hypothesis, it needs to be proven through evaluation.))
8.Work load Average/day, 9.**Hit target** (Work load is directly connected to Hit target which is also directly connected to absence frequency (Hit target would be directly correlated with many other features.))
10.Disciplinary failure (Actually, I tried googling about this term only to have failed...)
11.Education (Failure to graduate from high school can lead to the judgment that there is a problem with diligence, which affects absence frequency.)
12.Children, 13.Social drinker, 14.Social smoker, 15.Weight, 16.Height, 17.Body mass index (All of these could be blended into certain physical strength and health condition index, which have impact on absence/attendance frequency(15.+16. =>17.)
18.Pet (Whether having pets or not could be index of judging emotional stability, which has correlation with diligence. (total 18)

2) In this case, I think a Decision tree model works 'slightly' better than a Random Forest model, for the gap between two models in terms of Acc and AUC each is not that large but their Runtime is pretty different each other.

```
--ML Model Output--

Decision Tree Acc: 0.61 (+/- 0.14)
Decision Tree AUC: 0.59 (+/- 0.13)
CV Runtime: 0.03788566589355469
Random Forest Acc: 0.68 (+/- 0.09)
Random Forest AUC: 0.65 (+/- 0.11)
CV Runtime: 1.033799409866333
>>>
```

We should take the case that we conduct this learning with twice, a hundred times bigger data in consideration. The more data, the bigger runtime gap. When we use ML and judge how suitable it is for business, how long it does take does matter.

3) 6 features are selected - 1.Month of absence, 2.Day of the week, 3.Season 4.Transportation expense, 5.Work load Average/day, 6.Hit target

Feature Selection is about the selection of characteristics that are 'most useful' for training. We're trying to figure out which features can give us insight into the absence frequency. So thinking of why people are absent, we should consider workplace-related factors and external factors. In terms of internal and external factors, people will be absent because they are tired or lazy, and will be absent because they don't want to work due to hardship. The more expensive transportation cost, the more reluctant we become to pay it, which means that the worker may not want to transport (=**No Work**!) the *Now, It is clear why the six features above were selected*.

4)

```
--ML Model Output--

Decision Tree Acc: 0.54 (+/- 0.16)
Decision Tree AUC: 0.51 (+/- 0.18)
CV Runtime: 0.026900768280029297
Random Forest Acc: 0.55 (+/- 0.24)
Random Forest AUC: 0.49 (+/- 0.26)
CV Runtime: 1.0128285884857178
>>>
```

Those run-times are apparently different, but the difference between them is not that large; insignificant. 0.02 second and 1 second are considered

just as same in real world because both are very short time so the gap between them could be ignored. But our given dataset has only 700 rows ish. If the number of rows grows up to 70 million, a Random Forest model would take pretty much time to yield result. As the old saying, 'Time is gold', in business not only accuracy matters but also run-time matters. The fact that it takes much time for a Random Forest model to operate could be the significant problem in itself.

5) I'll suggest my boss that he reduce work load or provide incentives to workers who show high performance in spite of demanding work.

Fundamental manipulation any of 4 selected features except for 'Work load' is almost impossible - my boss cannot change the season according to workers' taste and lower the cost of transportation (boss could support transportation costs, but anyway the fact that transportation cost in itself is expensive is a problem.) and or hit target (it is possible but the company's income would directly damaged!). So we'd better mainly focus on the feature 'Work load', using what given closer examination showed, and can improve the workers' efficiency by simply reducing work load or compensate for their overwork.

6) The second model is preferable than first one. Seeing the features of Model #2, it seems it has 3 more features and higher accuracy by 0.8% than the first one so the second model is better one. And especially the features ‘BMI’ and ‘Education’ are very unique – ‘Hit target’ turned out to be correlated with ‘Work load’ feature but ‘BMI’ and ‘Education’ were unexpected because seemingly they have no correlation with given four features. But they came out through feature selection, which means implicitly they do have correlation with four features. In a conclusion, Model #2 has more consider-worthy features and higher accuracy so it is better than the first one.

7) We conjugate feature engineering to make data analysis easier when having hardship in using raw data. So I will mention ‘Why given raw data are hard to use’ first. (I will thoroughly consider given raw data from now on.)

When we compare Person1 with Person3 in health state, it is clear that Person2 has better health than Person1 because P2's (Smoke, Obese) is (1,1), and P2's is (0,0) - they are comparable to each other. But what if the situation is (1,0) versus (0,1) like the case of P2 and P6? Which one is more hazardous? To figure it out, we should weigh the hazard by scaling so feature engineering as well.
Also, I think 'Smoke' feature is insufficient in itself so had better alter into another or other feature(s) (of course obese as well). Let me compare P1 with P5. Are they in completely same health state? If not, who's smoking is worse for heath? We cannot know how severe their health became due to smoking. Also, Smoking could lead to way too various diseases (including Obese) . So it is better to divide the feature 'Smoke' into Smoking-related diseases such as tumor and cancer.

8) Without any changes such as feature engineering, My ML model for Korean customers would absolutely not suitable for American market. When I build a ML model to predict which cars customers are likely to buy in Korea, there would be features which are uniquely applied based on Korea such as 'The distance from customer's residence to agency. The data value range is from 0 to about 1000km in Korea, while it is from 0 to 2680 miles(about 4300km) which is extremely larger than maximum feature value in Korea. Because the feature previously I utilized is for Korea but not for America, we should use the Scaling method, one of the feature engineering, to re-size the value range so that my model could be usable even for American customers.

9) Seemingly the same case with problem 8), but I think it is quite different. Thinking about the principles of image recognition, it consists of processes of looking for a particular shape, line and pattern because those factors aren't different from country to country, and they're all in the same standard. In addition, the shape of airplanes are roughly the same in general so I think there would be no severe problem.

10) Absolutely the fist one is preferable. The problem found in Model 1 is typical so-called overfitting. The second model draws line quite smoothly while the first one in a tortuous and curvy way. This means that Model 2 is likely to find patterns in the training set which are not present in the

general data set just as ‘now’. But our boss must not be expert on data science so is going to have difficulty understanding the terminology ‘overfitting’. So I’ll explain like this:

> “Boss, you could think that the model 1 is better than 2 because seemingly it classifies much accurately in this case. But what if another data set is given? And another again? And countless others again? Could model1 still patternize and generalize for the other novel sets? We should keep in mind that this process is operated *not only once*!”

11)
column X : (the average, standard deviation)

column A: (0.367568, 0.482143)
B: (6.324324, 3.433964)
C: (3.914865, 1.420714)
D: (2,544595, 1.11108)
E: (2221.3297, 66.90697)
F: (29.63108, 14.82676)
G: (12.55405, 4.38191)
H: (36.45, 6.474393)
I: (271.4902, 39.03172)
J: (94.58784, 3.776759)
K: (0.054054, 0.226124)
L: (1.291892, 0.672783)
M: (1.018919, 1.097747)
N: (0.567568, 0.495414)
O: (0.072973, 0.260092)
P: (0.745946, 1.317367)
Q: (79.03514, 12.8745)
R: (172.1149, 6.030915)
S: (26.67703, 4.282556)

12)
The correlation between Transportation Expense and Distance from Residence to Work : 0.262183

The correlation between Age and Body Mass Index : 0.470688

I did expect the correlation of a. would be higher than that of b. because it seems Age and BMI are proportionate, with each value of each feature itself small, with little variation for me, and my expectation corresponded to the result of excel formula. But actually my prediction was more of a hunch because I couldn't calculate the correlation value of such a great deal of data - I underwent hardship even calculating only from row2 to row10 (even this was eye measurement without calculating!).

13)
The correlation value between 'Service Time' and 'Hit Target' is -0.00784, which means that the two features are inversely proportionate but that they have little to do each other because the absolute value of correlation is very close to 0. So I'll go to my boss and say like this, without any hesitation.

> "It is considerate of you to donate to workers who have long been our workplace although those bonuses have no improving effect on performance unlike your intention so,, I'll never invest in your program" (*talking in much more courteous way when really doing so :D)

14) The result from the t-test for Transportation Expense : 7.21507E-11 (A)

The result from the t-test for Age : 0.434927934 (B)

A is significant because it is lower than 0.05 but B is insignificant because it is higher than 0.05. So it is simple to explain the results to my boss.

> "Boss, We conducted so-called t-test that determines whether the feature is significant for absence frequency analysis. If the t-test value is higher than 0.05, it means that the feature is almost meaningless in analysis. As a result, Transportation Expense turns out to be fairly influential on absence frequency while Age turns out to be not because its t-test value was much higher than 0.05.

15) Classification and regression are clearly incomparable because the way each of them is assessed is totally different. Classification is to predict class which is already fixed, and in this situation the classification is binary, so class should be divided into two classes. On the other hand, regressions are usually used to solve problems in which prediction value is in the form of float, consecutive numbers, not a class. Thus evaluation is processed in different way. Accuracy, precision, recall, F1 score are evaluation metrics for classification and SSE or R2 is for regression.

16) In my opinion, rather than saying "It is accurate" or "It is not accurate" , I'd better say "I cannot judge". It is natural that the ML model that predict which computers will malfunction has such high accuracy of 89 percent, for the two class 'having malfunction' and 'having no malfunction' which are subjects of prediction were biased to one side - 10% versus 90% is undisputedly one-sided game! This fact that in the first place one class is in dominant position than another one might be a "red flag". To prevent this red flag, we should re-sample imbalance data so that data are properly distributed.