Identification of Cancer Essential Genes using GAMs

Abstract

In machine learning often a tradeoff must be made between accuracy and intelligibility. More accurate models such as boosted trees, random forests, and neural nets usually are not intelligible, but more intelligible models such as logistic regression, naive-Bayes, and single decision trees often have significantly worse accuracy.

XAI has various applications in the medical industry as if one questions ‘why this gene has been classified as essential, how are you sure that removal of this gene will mitigate cancer cells?’, the availability of an explanation becomes valuable and the users have a higher chance of accepting the answer given by the model.

Also if we have the answer to the question, “based on what features, does the model predict if a gene is essential for a cancer cell”, we will have a direction for the future research on the same. Considering only the features which significantly affect the health of a cancer cell. It can also be used to provide some counterexamples on previously thought important factors affecting the health of the cell.

Classes

The dataset comprised of two classes, whether the gene is essential or not. Our task is a binary classification problem.

Dataset

The dataset is divided, in the following parts

The input feature contains-

The frequency of 17 types of mutations that can happen to the gene (eg Missense Mutation, Silent).

The frequency of the type of Substitution happening for the particular gene (eg A->G, C->T).

The minimum and the maximum expression value the gene exhibits across the different mutations.

The frequency of the Copy Number the gene exhibits across different mutations.

Pre-Processing

The available data was processed to get the features in the following format.

The type of mutations in a gene was divided by the total number of mutations for that gene to get the fraction of the type of mutation in the gene.

The fraction of the type of substitution observed in a particular gene.

The minimum and the maximum value of the expression value observed of that gene in the mutated cancer cells.

The fraction of the value of the copy number observed in a particular gene.

Training Procedure

A Logistic GAM model is trained on the feature space of 37. GAM has a higher variance than Linear Regression as it allows non linear contribution of the features.

$g(E(y|X)) = f1(X1) + f2(X2) \ldots.. + fn(Xn)$

In the case of Linear regression, fi and g are Identity functions.g is known as the link function.

There is a link function which is chosen as logit, as it is the standard used for logistic regression.

Lam is a hyperparameter which varies the extent of regularization and helps to smoothen the function.

To decide on getting good value of lam for the model, randomly 100 samples of lam were generated between (exp(-3),exp(3)) and the model with the lowest generalized cross-validation score was considered.

The partial dependence plot was plotted for the various features to get an idea about the variation of the output with the feature.

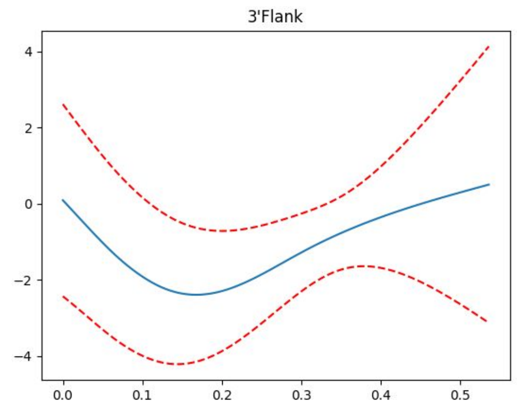
Results

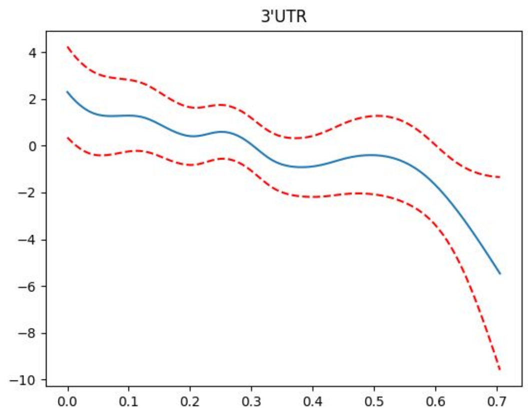
Accuracy = 71%

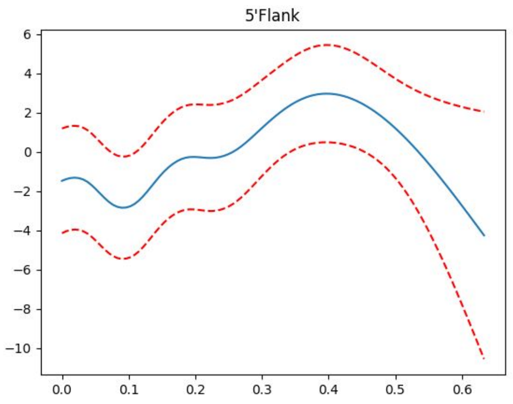
To evaluate how the different features affect the model, we plot the Partial Dependence Plots.

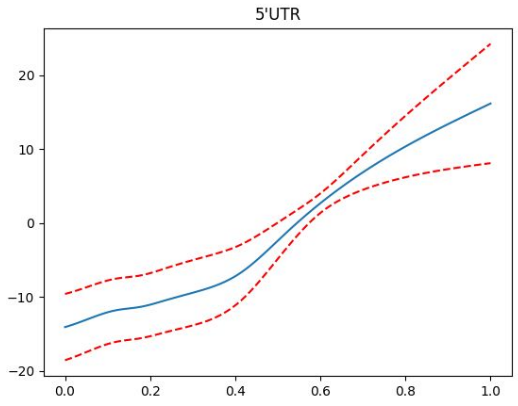
Partial Dependence Plot

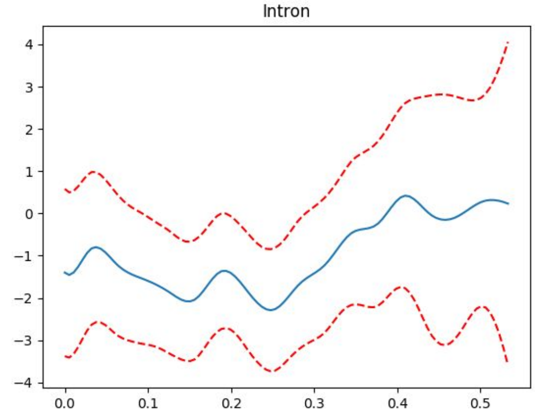
1) Mutation Types

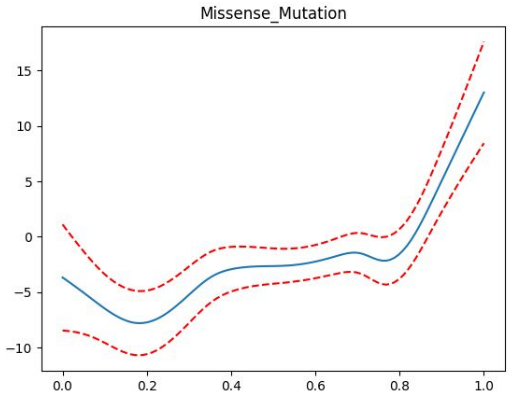


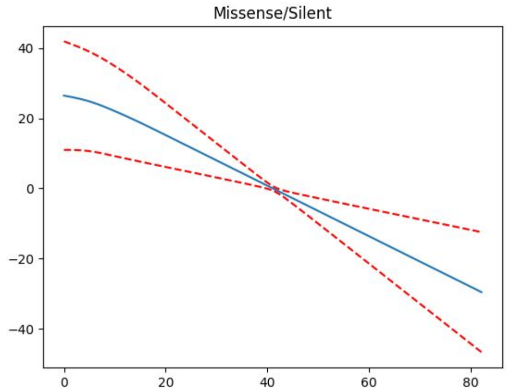


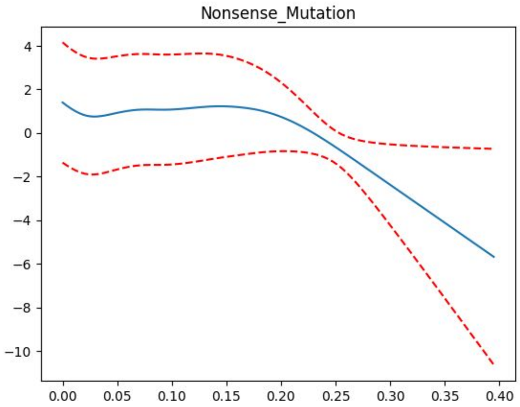


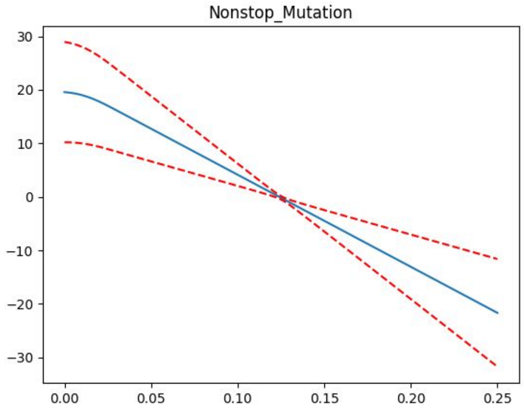


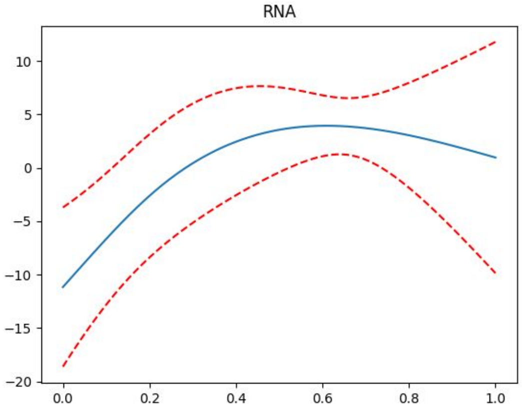

Intron
4
3
2
1
0
−1
−2
−3
−4
0.0
0.1
0.2
0.3
0.4
0.5

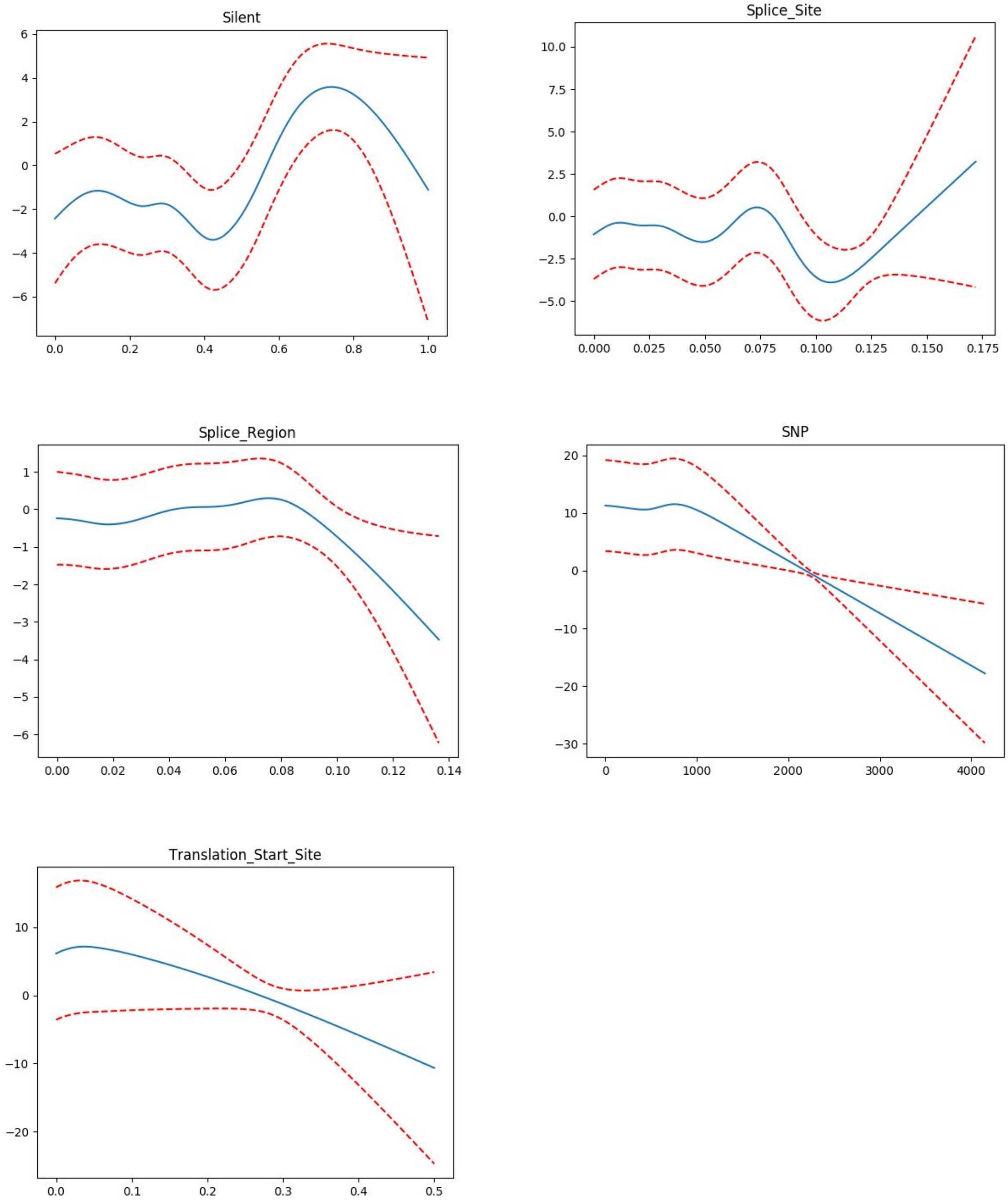

Missense_Mutation
15
10
5
0
−5
−10
0.0
0.2
0.4
0.6
0.8
1.0


Missense/Silent
40
20
0
−20
−40
0
20
40
60
80


Nonsense_Mutation
4
2
0
−2
−4
−6
−8
−10
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40


Nonstop_Mutation
30
20
10
0
−10
−20
−30
0.00
0.05
0.10
0.15
0.20
0.25


RNA
10
5
0
−5
−10
−15
−20
0.0
0.2
0.4
0.6
0.8
1.0



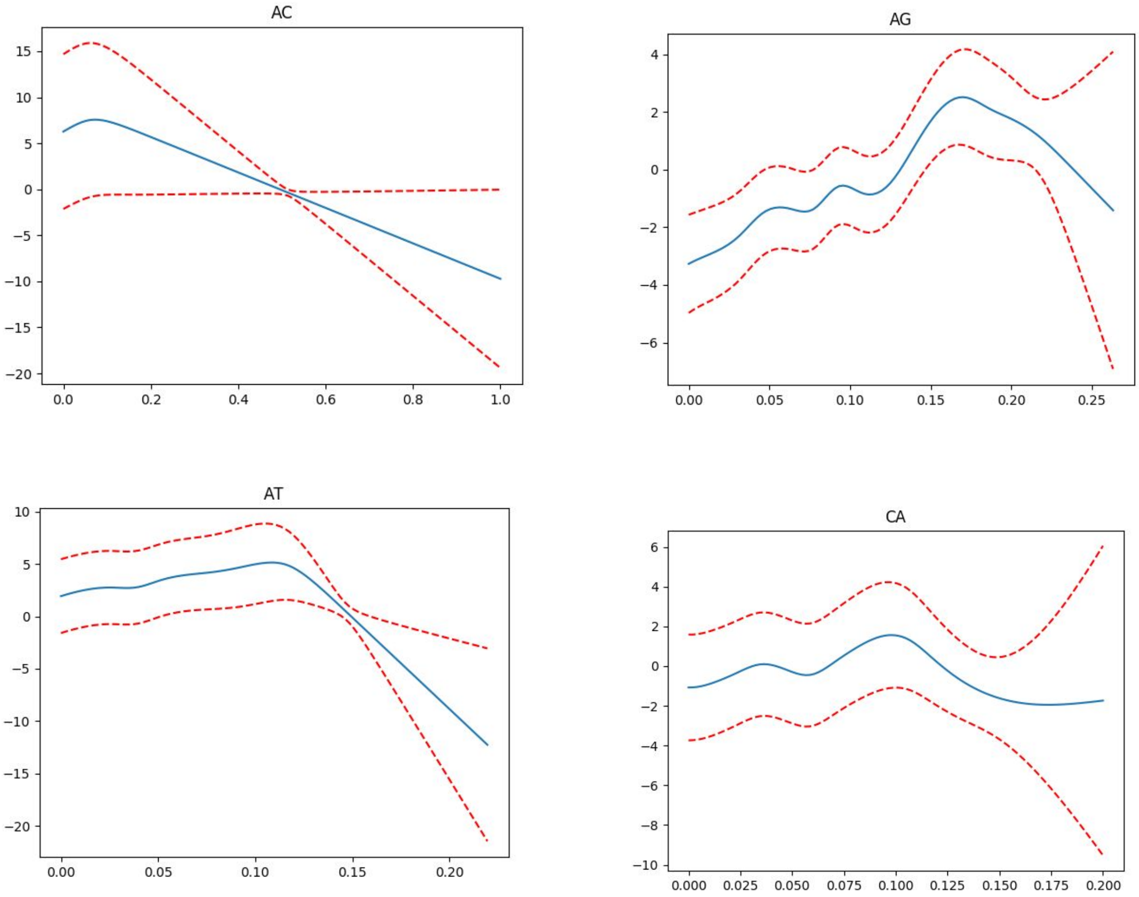
The blue line represents the expected value of g(y) and the red lines represent the boundary of 95% confidence. I:e g(y) will lie in this range with a probability of 95%.

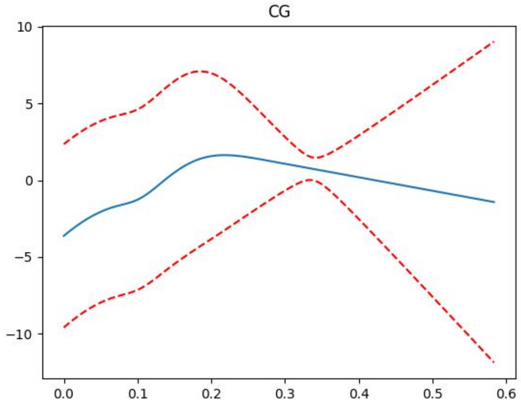
In the case of MIssense/Silent mutation ratio, we observe that the probability of the gene being classified as essential increases this makes sense as the mutation of an essential gene would generally not lead to a silent mutation.

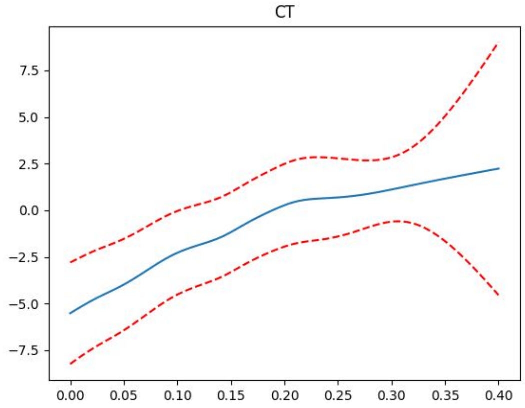
SNP (Substitution Mutations) represents the total number of mutations of the gene in the dataset, as SNP increases it is expected that the gene is not essential as we don't expect an essential gene to undergo mutation as it is required by the cell for it's functioning.

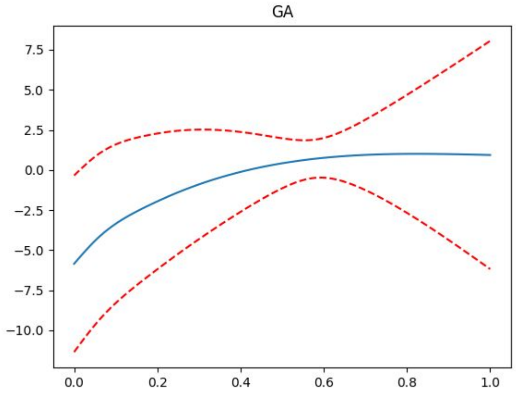
2) Nucleotide Change

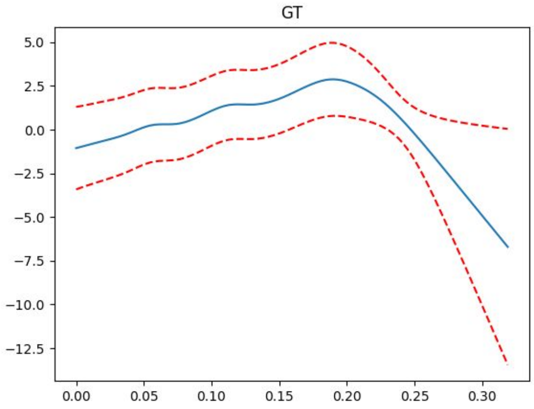
The following plots represent the variation of the type of nucleotide change with the output. Here, for example, AT represents A nucleotide is replaced by T nucleotide.

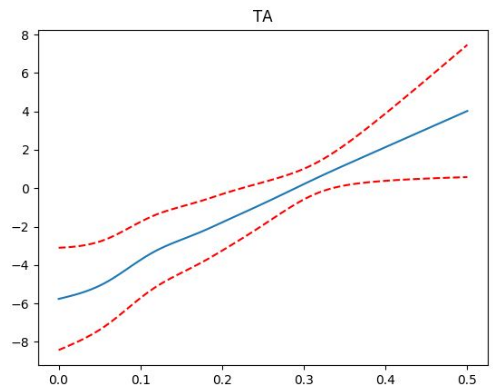


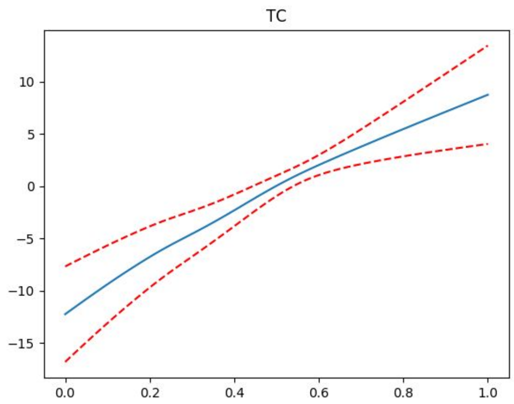

CG
10
5
0
−5
−10
0.0
0.1
0.2
0.3
0.4
0.5
0.6

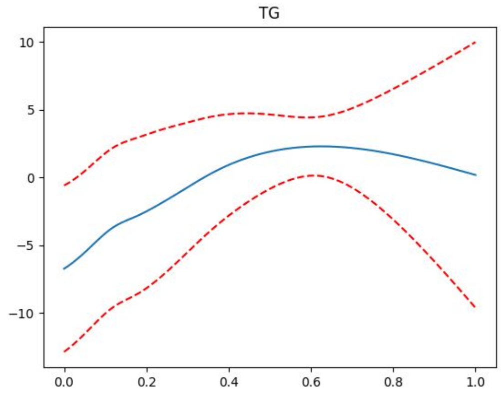

CT
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40


GA
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
0.0
0.2
0.4
0.6
0.8
1.0


GT
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
0.00
0.05
0.10
0.15
0.20
0.25
0.30



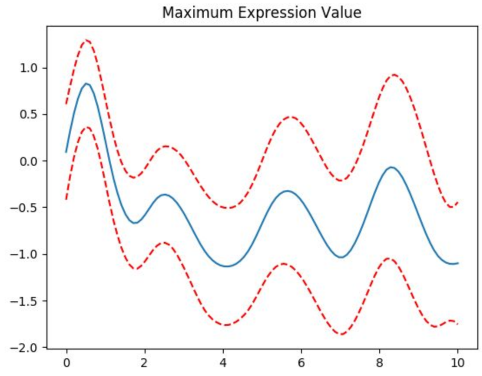


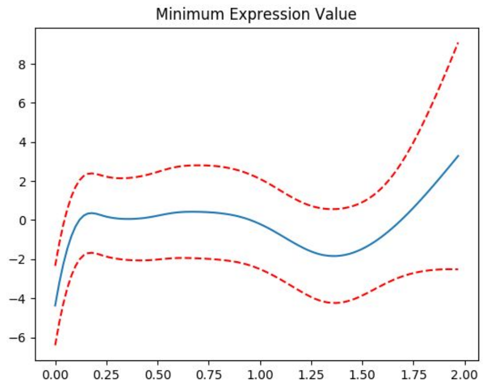


The plots indicate that if it is more likely that a gene will undergo a T->C, T->A or C->T substitution, it has a higher expectation of being an essential gene.

While on the other hand if the substitution of type A->C or A->T has a higher chance of taking place, the gene is more likely to be a non-essential gene.

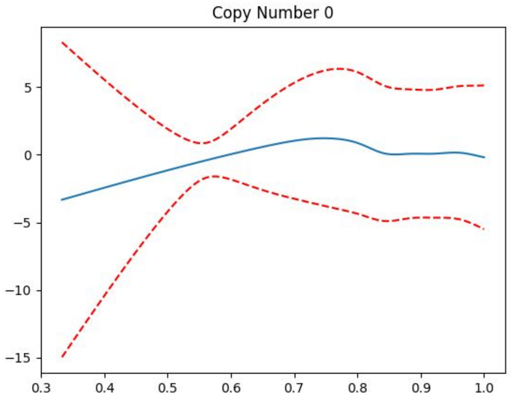
3) Expression Value

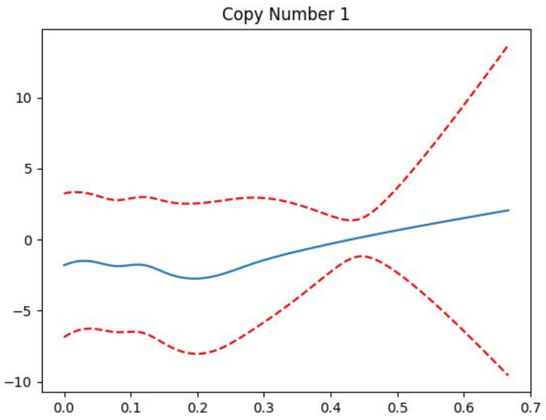


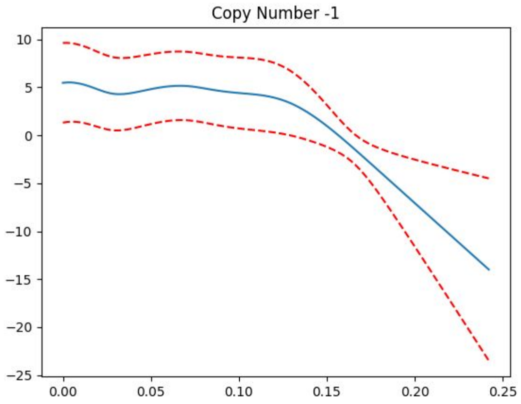


Here we observe that at a lower value of the gene expression data, it is more likely for a gene to be non-essential as it is expected from an essential gene to have a higher expression value as it is a representative of how active is the gene in the cell. The maximum expression value does not seem to give a proper model of variance.

4) Copy Number fraction







Conclusion

The model seems utilised GAMs to help us discover previously unknown relations between certain features and the probability of a gene being essential, for example, a gene which has a higher probability of undergoing an A->T transition is more likely to be a non-essential gene. It also shows some expected trends like increase in the Minimum Expression value of a gene means a higher probability of the gene is an essential gene.

So, in conclusion, though models with higher variance like Neural Nets might provide better accuracy to study the trends of the effect of different features more transparent models like GAM work well with decent correctness.