Predict who would be interested in buying a Caravan Insurance Policy

For Assignment 3, we will use The Insurance Company Benchmark (COIL 2000) dataset. You can access the data from following link.

http://kdd.ics.uci.edu/databases/tic/tic.html

This data set used in the COIL 2000 Challenge contains information on customers of an insurance company. The data consists of 86 features and includes product usage data and socio-demographic data derived from zip area codes. The data was collected to answer the following question: Can you predict who would be interested in buying a caravan insurance policy and give an explanation why?

Three data files are available at given link. **Ticdata2000.txt** contains training data, 86th column is target value. **Ticeval2000.txt** is test dataset and **ticgts2000.txt** is target value for test data. All the description required for this dataset are available at given link in the beginning.

Wherever needed, write down your observation in comments in your notebook. **Best model would be considered based on its f-score on both classes.**

1. Which features are relevant for the prediction task? Select top 10 features based on your understanding. Show visualizations or statistics to support your selection. [10 Marks]

2. Train a Logistic Regression (LogReg) model with L1 regularization. Find the best model using grid search on **C** values. Analyze which features have nonzero coefficients for the best model. Are they in synch with your selected features from question 1? [20 Marks]

3. Generate polynomial features and use LogReg again with L1. See if accuracy increase. [20 Marks]

4. Use any classification model we discussed (trees, forests, gradient boosting, SVM) to improve your result. You can (and probably should) change your preprocessing and feature engineering to be suitable for the model. You are not required to try all of these models. Tune parameters as appropriate. [30 Marks]

5. Can you create an "explainable" model that is nearly as good as your best model? An explainable model should be small enough to be easily inspected - say a linear model with few enough (<10) coefficients that you can reasonable look at all of them, or a tree with a small number of leaves, less depth etc. [20 Marks]