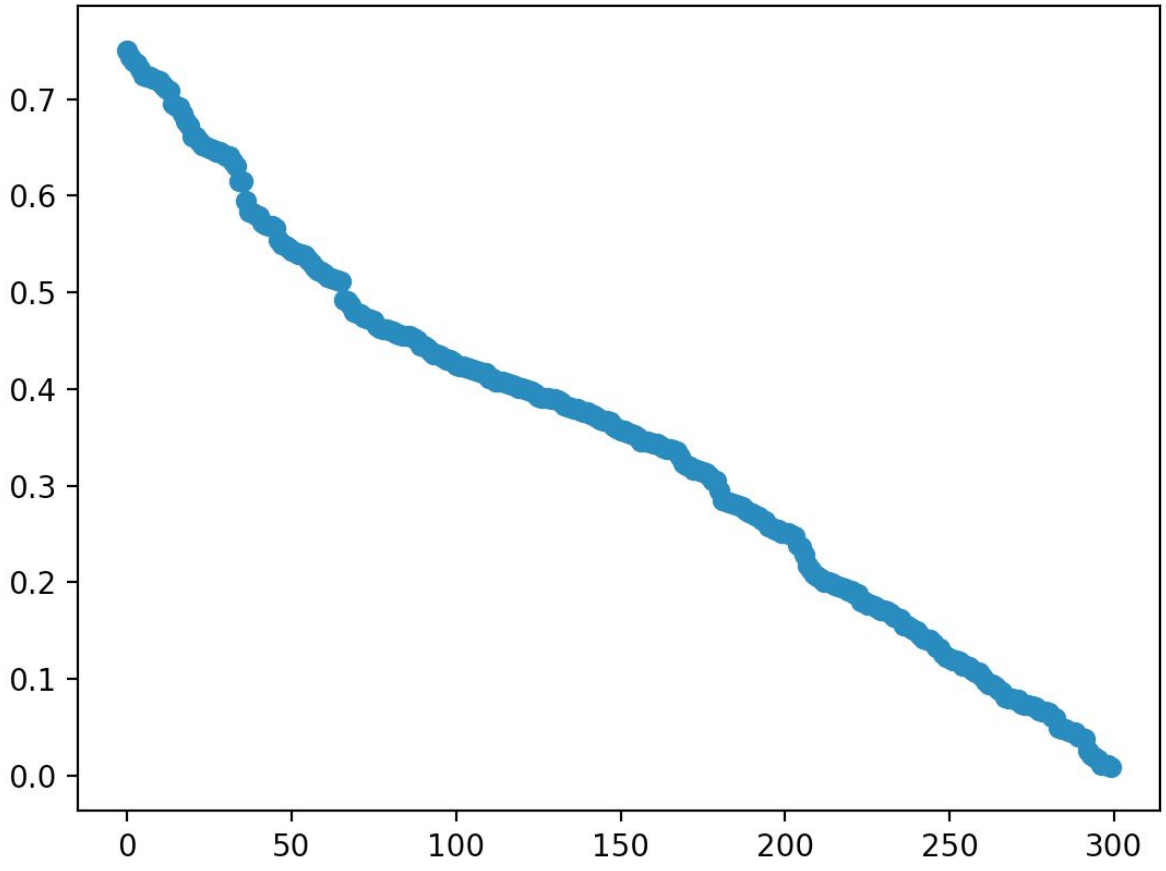
Direct Mail Prospect Scoring Exercise

Ques 1. Logistic Regression equation:
logit(y) = -10.36654812 + Gender*-0.04724382815367635 + foodwebsites*0.005967340175320205 + travelwebsites*0.04386315715791136 + restaurantexp*0.06982632349984076 + travelexp*0.03414963790885262 + entertainmentexp*0.008917756290912748 + incm*-0.0009067094634970179 + ethnicdivneigh*-0.0008393870374434614

Ques 4: curve for marginal response rate vs number of solicitations made

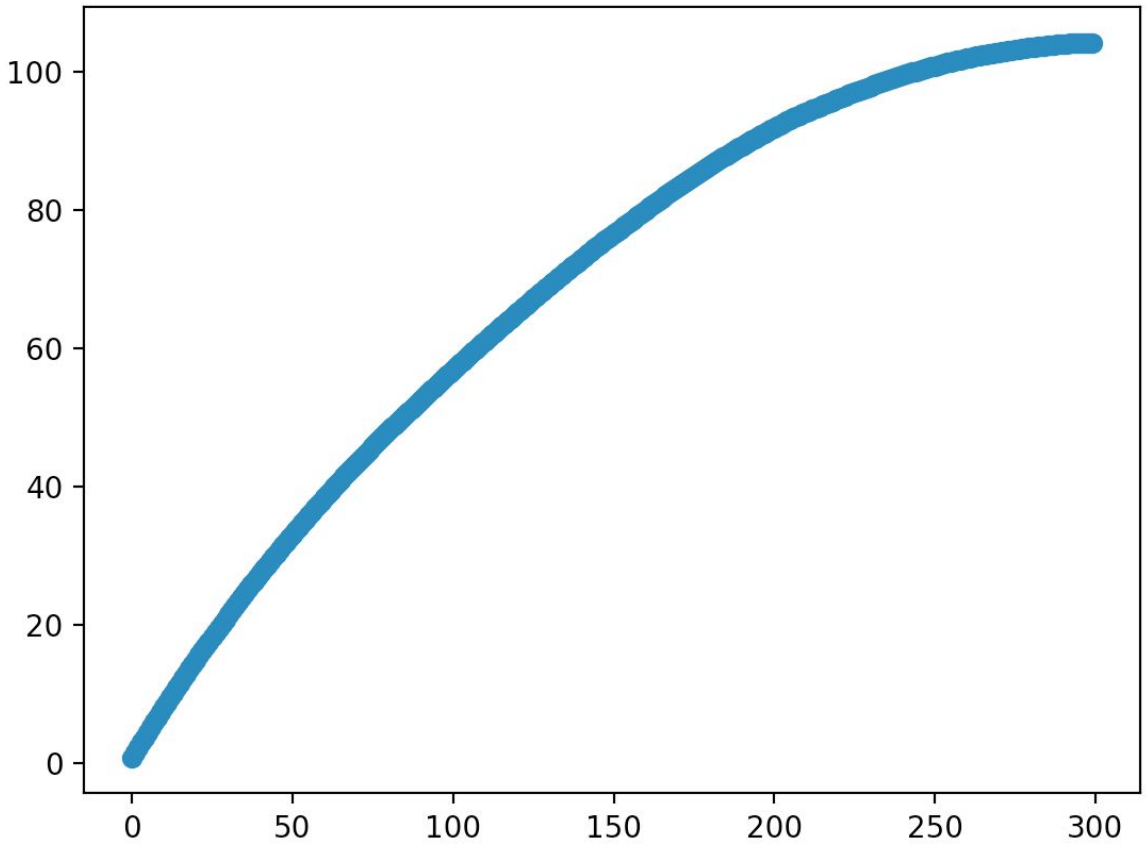


Ques 5:

121

Ques 6:

There is no huge decrease due to a less variation in the data. It seems as though, the data is more distributed as compared to the real world.

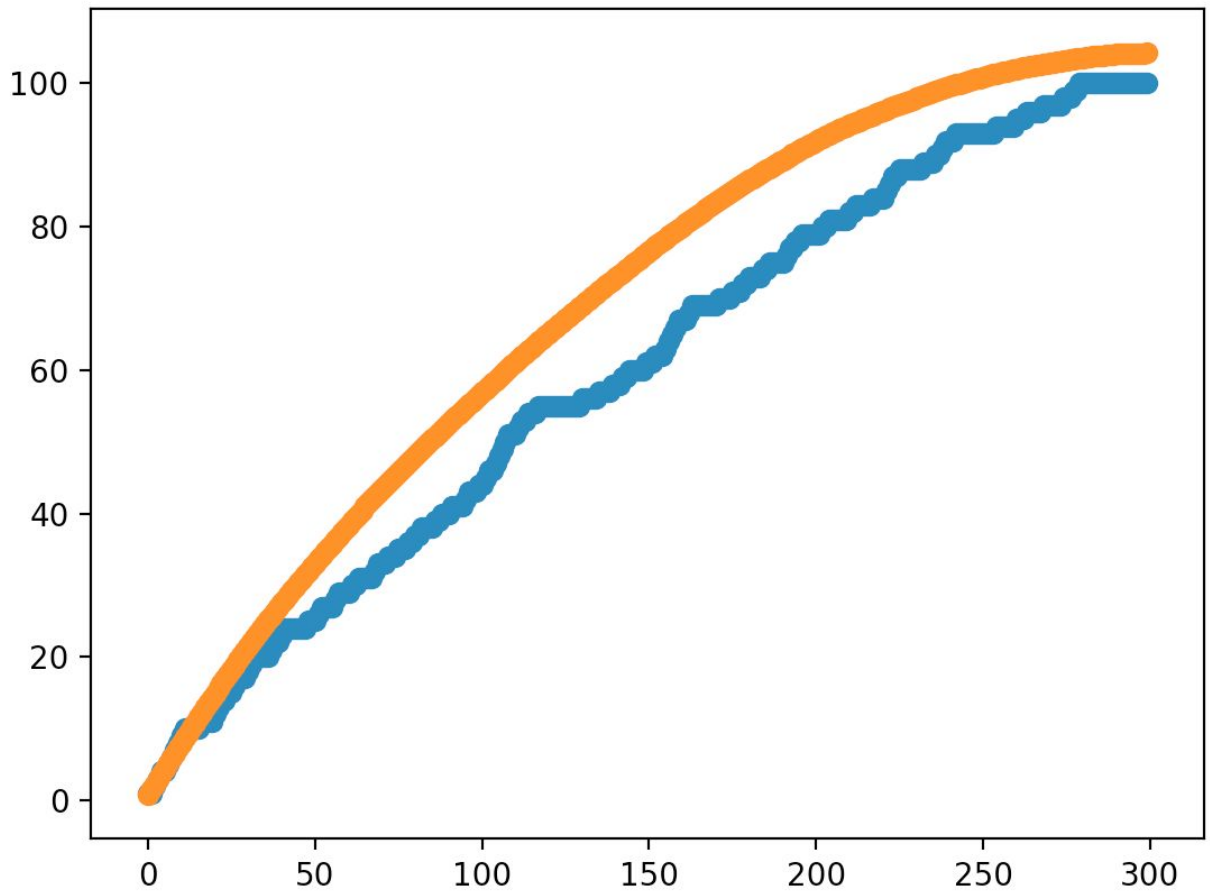


We do not see an exponential rise or the pareto principle followed in our graph due to 2 main reasons. The first being the assumption the pareto principle follows. The 80-20 distribution is only possible if the graph's distribution has a long tail. If the data is following is a gradual decrease (as shown in question 4) then the 80-20 distribution would not be followed. This is easy to understand. The second reason why the principle doesn't hold is because of the variance of our data. When we perform logistic regression and deduce the probability, we perform predictions using the maximum likelihood estimator (the goal being the most optimal way to fit a distribution to our data). In doing so, we select that estimator which has the best likelihood to our dataset, and due to this in our dataset we don't see any values with an extremely high / low probability. This in turn leads to a graph like this. If our dataset was more scattered we might have been able to achieve the 80-20 distribution.

Ques 7:

Melrose should send solicitations to 64 of the best prospects in the holdout set.

Ques 8:



In our previous answer we had suggested sending solicitations to 64 of the best prospects. From our graph, we can deduce that most of them would opt in.

Question 9: