We investigate statistical aspects of subsampling for large-scale linear regression under label budget constraints. In many applications, we have access to large datasets (such as healthcare records, database of building profiles, and visual stimuli), but the corresponding labels (such as customer satisfaction, energy usage, and brain response, respectively) are hard to obtain. We derive computationally feasible and near minimax optimal subsampling strategies for both with and without replacement settings for prediction as well as estimation of the regression coefficients. Experiments on both synthetic and real-world data confirm the effectiveness of our subsampling algorithm for small label budgets, in comparison to popular competitors such as uniform sampling, leverage score sampling and greedy methods.