Probabilistic Expense Prediction

Problem Statement

According to Federal Reserve Surveys in 2017, 40 percent of U.S. households expected having trouble to pay for a $400 (emergency) bill[i]. While unexpected household expenses could result in unexpected financial, credit, and social costs, overestimating expenses can result in losing investment opportunity loss for households.

Can you use Machine Learning to estimate, as precisely as possible, the uncertainty distribution of a household's total monthly expenses?

Notice that point estimation of household expenses is misleading because it doesn't reflect the uncertainty associated with expenses (e.g. emergency and non-discretionary costs). For this reason, this case study is focused on estimation uncertainty as well as expected expenses.

Dataset

You are provided with a dataset based on the Consumer Expenditure Survey[ii]. The input features include consumer id, year, month, education level, age of the reference person, family size, urban residency, race, region of residency, state of residency, marital status, occupation, annual income (income), and consumer unit weight. The consumer unit weight shows how much a row is representative of US population. The target variable is the monthly expense column.

Assume that occupation, marital status, state, region, race, urban are categorical variables, education is an ordinal variable.

Notice that multiple age, education, race, and income are individual features of the reference person in a potential household while total expenses is a household-level variable.

The dataset is a robust and can be used to improve forecasting accuracy for individual households. The methods used in this project, if successful, can be applied to various business problems including in financial, retail, and manufacturing sectors.

Submission

The first column includes the consumer ids of test cases and the next 9 columns are quantiles: q_1=0.005, q_2=0.025, q_3=0.165, q_4=0.25, q_5=0.5, q_6=0.75, q_7=0.835, q_8=0.975, and q_9=0.995 corresponding to median and 67%, 75%, 95%, and 99% prediction intervals for a test case.

The submission file should be a table in csv format (test_quantiles.csv) where rows include customer ids from the test file, followed by 9 columns for expense prediction quantiles. For instance:

id	q1	q2	q3	q4	q5	q6	q7	q8	q9
741875	960	1563	2313	2540	3500	4427	4670	5450	6187
741876	1635	2189	2764	2956	3714	4356	4638	5021	5581
741877	1067	1551	2019	2286	2896	3501	3654	4225	4703

In addition to above result file, the training and test codes are expected in the submission package.

Evaluation Metric

The precision of the quantiles will be evaluated based on Pinball Loss function[iii]. The measure is calculated for each test case and quantile as follows:

$$\frac{1}{Nt * Nq}\sum_{i=1}^{Nt} w_i \sum_{j=1}^{Nq}\big(E_i - Q_i(q_j)\big)q_j \mathbf{1}\{Q_i(q_j) \le E_i\} + (Q_i(q_j) - E_i)(1-q_j)\mathbf{1}\{Q_i(q_j) > E_i\}$$

Where i is index for a consumer in test case, j is index for a quantiles (), Nt is the number of test cases, Nq is the number of quantiles (9), w_i is consumer unit weight, E_i is an actual expense for household i, $Q_i(q_j)$ is the expense prediction for quantile q_j, and $\mathbf{1}\{.\}$ is the indicator function.

Notes:

- You may use macroeconomic indicators (e.g. inflation, unemployment rate, etc.) to improve the model performance.
- You may train Variational Auto-encoders to learn the expense distributions and extract quantiles
- Consumer unit weight is not available in test data but you may use it for aggregating data or model tuning purposes

---

[i] https://crr.bc.edu/wp-content/uploads/2019/07/IB_19-11.pdf

[ii] https://www.bls.gov/cex/

[iii] https://www.lokad.com/pinball-loss-function-definition