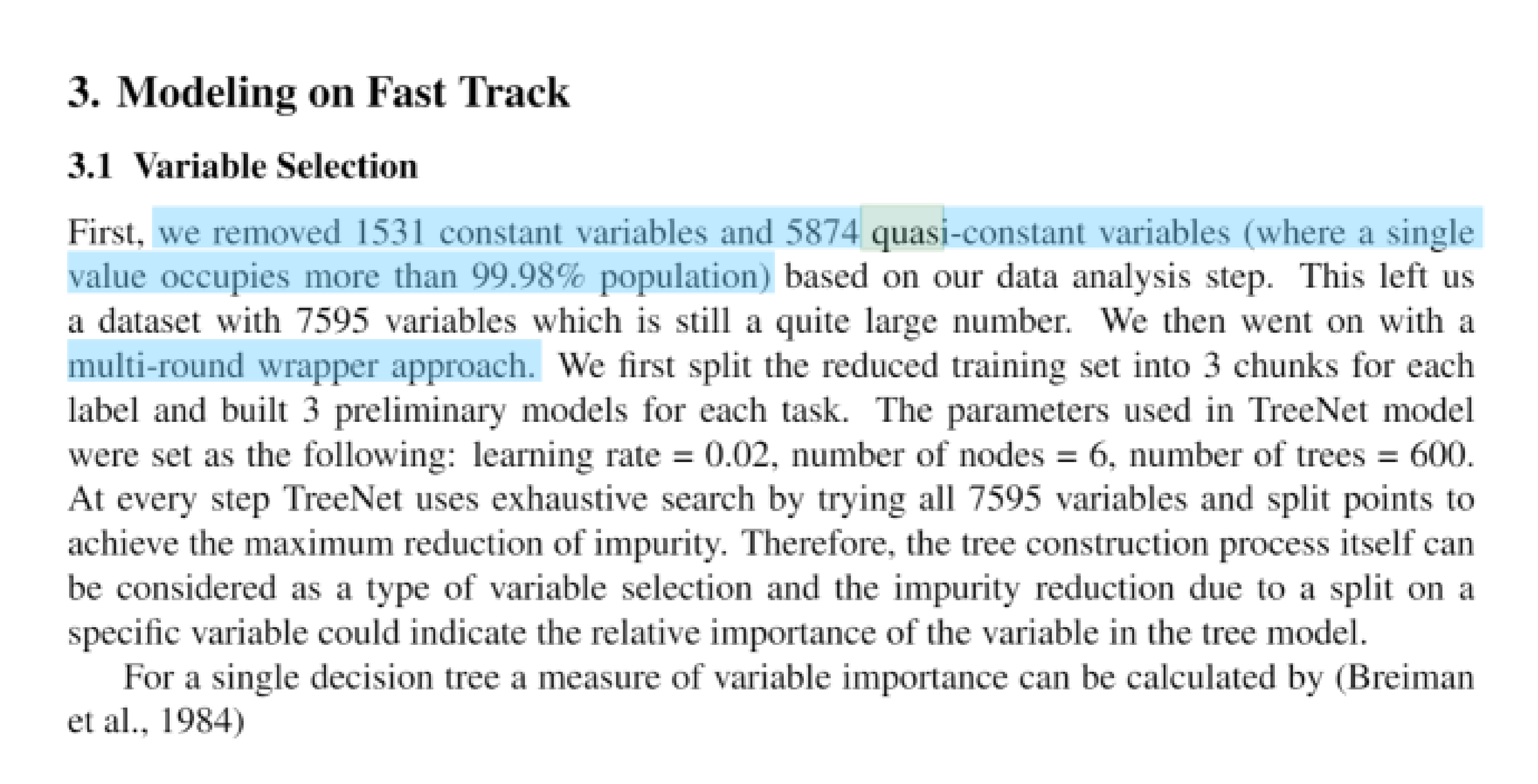
FILTER METHODS

BASIC METHODS

FILTER METHODS - BASIC METHODS

- Constant features
- Quasi-constant features
- Duplicated features
  - ➢ Duplicated features may arise after one hot encoding of categorical variables

FILTER METHODS - BASIC METHODS



3. Modeling on Fast Track

3.1 Variable Selection

First, we removed 1531 constant variables and 5874 quasi-constant variables (where a single value occupies more than 99.98% population) based on our data analysis step. This left us a dataset with 7595 variables which is still a quite large number. We then went on with a multi-round wrapper approach. We first split the reduced training set into 3 chunks for each label and built 3 preliminary models for each task. The parameters used in TreeNet model were set as the following: learning rate = 0.02, number of nodes = 6, number of trees = 600. At every step TreeNet uses exhaustive search by trying all 7595 variables and split points to achieve the maximum reduction of impurity. Therefore, the tree construction process itself can be considered as a type of variable selection and the impurity reduction due to a split on a specific variable could indicate the relative importance of the variable in the tree model.

For a single decision tree a measure of variable importance can be calculated by (Breiman et al., 1984)

A combination of boosting and bagging for KDD Cup 2009 – Fast Scoring on a Large Database. Xie, Rojkova, Pal, Coggeshall. JMLR Workshop and Conference Proceedings 7:35-43, 2009