Method

3.1 Data Collection

Dalam penelitian ini, kami menggunakan data dari PT. PLN Persero. Data yang diambil hanya 1 daerah karena field tiap daerah sama saja, data yang diambil PLN zona sumatera barat . Penelitian kami menggunakan data transaksi nasabah Januari 2019 hingga Desember 2020 yang terdiri dari 16.811.662 dan 90 variabel data dalam 2 tahun. Tabel 2 menunjukkan statistik deskriptif untuk data.Beberapa cantuman dihapus dari kumpulan data kaerna menunjukkan kesimpulan yang tidak logis yaitu catatan duplikat atau nilai yang hilang.

3.2 Choice Variable

Variabel prediktor didasarkan pada data yang diperoleh dari PLN Zona Sumatera Barat. Ada sekitar 90 variabel tetapi tidak semuanya digunakan karena variable akan ditentukan oleh dua model yaitu CRM dan KAM. Oleh karena itu, variable yang dipilih memiliki potensi yang tinggi dari 2 model tersebut . Tabel 3 dan 4 menunjukkan informasi rinci tentang variabel prediktor.

3.3 Choice Customer Segmentation

Pada penelitian ini bertujuan untuk menentukan segmentasi pelanggan dari variable yang sudah dipilih sebelumnya. Segmentasi pelanggan dibagi menjadi 3 yaitu Rumah ,Industri menengah , Industri Besar . Klasifikasi didasarkan pada variable CRM dan KAM yang telah dipilih sebelumnya yaitu berdasarkan daya yang dipakai, metode pembayaran, jamnyala yang dipakai oleh pelanggan dan lain- lain.Tabel 5 menunjukan dari segmentasi pelanggan

3.4 Choice Potential Method

Fokus pada penelitian adalah untuk mengembangkan model prediksi segmentasi pelanggan dari gabungan metode Customer Relationship Management (CRM) dan Key Account Marketing (KAM) dengan koefisien dan kesalahan standar yang dapat memprediksi secara akurat apakah kepuasan pelanggan mempengaruhi perusahaan. Tabel 4 menunjukkan model prediksi yang akan digunakan dalam penelitian ini. Melihat jumlah model bawaan dalam model pembelajaran mesin, peneliti menyelidiki model ensemble dan model tunggal. Secara umum, model ensemble lebih akurat dalam memprediksi daripada model tunggal [27]. Namun, model tunggal masih mengungguli model ensemble. Peneliti menggunakan Logistic Regression, Decision Tree, Random Forest Tree. Pada kelompok ensemble, peneliti menggunakan pendekatan random forest tree.

3.4.1. Logistic Regression

Generally, Logistic Regression is used to describe and test hypotheses [28]. Choosing the correct variable and avoiding highly correlated variables must be considered when using Logistic Regression [29]. The variable predictors in logistic regression can be categorical or numeric, and the target variable for linear regression is binary or dichotomous. Therefore, Logistic Regression cannot predict the target variable for more than two classes. Although Logistic Regression may have some disadvantages, it can often compete with other machine learning techniques, such as neural networks, machine support vectors, random forest, and gradient enhancement. Logistic regression formulation is stated as follows:

Where

3.4.2. Decision Tree

Decision Tree dapat memecahkan masalah klasifikasi. Seperti namanya, algoritma Decision Tree terlihat seperti struktur pohon. Ini memiliki akar simpul, simpul daun, dan cabang; dan beberapa keunggulan, seperti nonparametrik, adaptif dengan dataset apa pun, dan dapat berurusan dengan hubungan non-linear [30]. Decision Tree adalah algoritma yang digunakan dalam pohon keputusan [32] dan menggunakan indeks Gini untuk mengevaluasi perpecahan. Skor terbaik adalah 0, dan skor terburuk adalah nilai yang sama untuk setiap kelas. Formalisasi dari indeks Gini dinyatakan sebagai berikut [31]:

3.4.3. Random Forest Tree

Metode ensemble menggunakan Random Forest untuk pengklasifikasi, yang terdiri dari Pohon Keputusan yang dibentuk secara acak dan independen dari dataset sampel. Ini menggunakan hukum bilangan besar, sehingga tidak overfit dan bisa baik untuk prediksi [33]. Selanjutnya dapat digunakan untuk dataset apapun karena tidak memerlukan asumsi distribusi [34] tetapi kelemahannya adalah dapat menjadi bias karena sampel terdiri dari komposisi label prediksi yang berbeda [35]. Formalisasi pengklasifikasi hutan acak dinyatakan sebagai berikut [36]:

3.5. Evalution,Validation and Model Selection

To measure the prediction performance of the predetermined model, the researcher used two different evaluation methods, namely three-way split ten folds and ten-fold cross procedures. In a tenfold three-way data separation procedure, we perform two groupings of data. In the first grouping of data, the researcher divides the data set into ten equal parts or folds.

The dataset is split into ten folds and is not evenly divided. From 19,200,000 records, the researcher grouped the dataset to fold number one to number nine, which consisted of 192,000 records. The number tenfold consists of 193,000 notes. The second grouping is more functional. First, the training set is used to match the data points to the proposed model. Second, the validation set is used to evaluate the most accurate model trained in the training set. The third set of tests is used to generate the final predictive score for each generated model. The number of data records used in training, validation, and test sets was adjusted according to the fold number category. If the test is set to number ten (192,000 records), the training set consists of 1,544,000 (192,000 x 8 times) and the validation set consists of 192,000 records.

If the test is not set to number ten (192,000 records), the training set consists of 1,500,000 records (190,000 x 7-fold + 192,000 records from the previous ten-fold) and the validation set consists of 192,000 records. In total, there are 900 test combinations. In the second procedure, tenfold cross-validation, we split the data into ninefold for training and one fold for testing.

In total, there are ten test combinations. Predictive scores of the evaluated models using ten-fold and tenfold cross-validation procedures were compared. The model with the highest predictive score is selected. In this study, the receiver operating features (ROC) or simplified AUC value were used to determine the prediction score better than accuracy. Mathematically, the researchers formulated the ABK score as follows:

E. Model Use and Reporting

Mereka membandingkan waktu pengembangan model dan skor prediksi berdasarkan kinerja masing-masing model. Model prediktif terbaik dengan skor AUC-ROC prediktif terbaik akan membantu pengambil keputusan dalam merumuskan kombinasi CRM dan KAM yang tepat dengan cara yang lebih baik.