Loan Interest Rate Prediction

**Problem Statement:**

Have you ever wondered how lenders use various factors such as credit score, annual income, the loan amount approved, tenure, debt-to-income ratio etc. and select your interest rates?

The process, defined as ‘risk-based pricing’, uses a sophisticated algorithm that leverages different determining factors of a loan applicant. Selection of significant factors will help develop a prediction algorithm which can estimate loan interest rates based on clients’ information. On one hand, knowing the factors will help consumers and borrowers to increase their credit worthiness and place themselves in a better position to negotiate for getting a lower interest rate. On the other hand, this will help lending companies to get an immediate fixed interest rate estimation based on clients information. Here, your goal is to use a training dataset to predict the loan rate category (1 / 2 / 3) that will be assigned to each loan in our test set.

You can create feature engineering by using any combination of the features in the dataset to make your loan rate category predictions accuracy higher. Some features will be easier to use than others.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Definition** | |
| Loan\_ID | A unique id for the loan. | |
| Loan\_Amount\_Requested | The listed amount of the loan applied for by the borrower. | |
| Length\_Employed | Employment length in years | |
| Home\_Owner | The home ownership status provided by the borrower during registration. Values are: Rent, Own, Mortgage, Other. | |
| Annual\_Income | The annual income provided by the borrower during registration. | |
| Income\_Verified | Indicates if income was verified, not verified, or if the income source was verified | |
| Purpose\_Of\_Loan | A category provided by the borrower for the loan request. | |
| Debt\_To\_Income | A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower’s self-reported monthly income. | |
| Inquiries\_Last\_6Mo | The number of inquiries by creditors during the past 6 months. | |
| Months\_Since\_Deliquency | The number of months since the borrower's last delinquency. | |
| Number\_Open\_Accounts | The number of open credit lines in the borrower's credit file. | |
| Total\_Accounts | The total number of credit lines currently in the borrower's credit file | |
| Gender | Gender | |
| Interest\_Rate | Target Variable: Interest Rate category (1/2/3) of the loan application | |
|  | | |

|  |
| --- |
| Data Exploration   Data diinput Train data dan Test data, lalu Labeled data Interest Rate Category (1/2/3) pada train dijadikan variabel yang akan diprediksi.      Kemudian dilakukan Explorasi Data dengan melihat karakteristik data, summary data, null value  Mengecek data kosong dengan seaborn heat map. B) Data pre-processing Membuat function untuk menemukan data kosong dari mulai jumlah total, persentase, membuat table summary data kosong.      Kemudian menghapus data dengan nilai persentase lebih dari lebih dari 50% yaitu (Months\_Since\_Deliquency)    Load\_ID dihapus karena merupakan unik ID, tidak ada hubungan apapun dengan predictor.    Mengisi kolom yang masih memiliki data kosong yaitu Home\_Owner, Annual\_income, dan Length\_Employed. Kemudian Membersikan data pada kolom Length\_Employed dengan menghapus “year” “years”, dan tanda “+”, mengganti nilai <1 year dengan 0,5 dan >10 years dengan 10.   * Untuk Home\_Owner diisi dengan Mortgage karena merupakan data yang paling sering muncul      * Untuk Annual\_income Diisi dengan nilai rata-rata      * Untuk Length\_Employed diisi dengan rata-rata lama kerja dengan asumsi lama bekerja >10 tahun dianggap 10 tahun.   Setelah lengkap, data type dicek.    kolom yang bertipe Object diganti sesuai dengan value pada kolom masing masing. Karena akan diubah menjadi numerical value untuk beberapa kolom yang mengandung string seperti Gender dengan 0 atau 1    Di akhir sebelum melakukan pemodelan dan prediksi. Data dengan value string di numerical kan dengan Sklearn Preprocessing Label Encoder agar semua data menjadi matrix angka    Sebelum Melakukan Pemodelan data Train dibagi menjadi 80% sebagai trian dan 20% menjadi test dengan sklearn.model\_selection guna melakukan penentuan evaluasi/accuracy dari beberapa model yang dipilih.   C) Modeling |
| Pada tahap pemodelan dilakukan dengan dua jenis Classifications Model yaitu DecissionTreeClassifier dan RandomForestClassifier.    Untuk model DecissionTreeClassifier didapat accuracy sebesar 50,3%  Untuk model RandomForestClassifier didapat accuracy sebesar 51,7%  Sehingga model yang dipilih untuk penentuan prediksi Interes Rate pada data Test yaitu Random Forest.  Kemudian diexport menjadi data Excel. |