1. The fields you selected: 'Vintage','ChargeOffMOB','BalanceAtDefault','FICOScore','FICOScorePctAvgFICOLast30DaysBookedLoans','NewUsedIndicator','VehicleManufacturerRebate','VehicleValueBlackBook','MoodysGDPReal','FinancedAmt','CashDownPmt','CashDownPmtPctBBVehicleValue','PaymentAmt','PaymentToFinancedAmtRatio','PTILine5','BackendTotalAmt','TotalDownPmt'

We selected these fields as they are the deciding factors to come up with a recovery rate for an auctioned car.

Vintage: signifies the timeframe of inception of account

'ChargeOffMOB': start charge off loan

BalanceAtDefault: Total balance defaulted

FICOScore: credit score while applying loan

FICOScorePctAvgFICOLast30DaysBookedLoans: FICO score as a percentage of "prior" rolling 30 day

'NewUsedIndicator': Whether vehicle is new or used

'VehicleManufacturerRebate': whether cast rebate applied in loan

'VehicleValueBlackBook': Vehicle auctioned value

'MoodysGDPReal': GDP value at mentioned period

'FinancedAmt': Amount disbursed to the loan

'CashDownPmt': Cash down payment

'CashDownPmtPctBBVehicleValue': Cash down as % of auctioned value

'PaymentToFinancedAmtRatio': Monthly payment to financed amount

'PTILine5': Payment to income ratio

'BackendTotalAmt': Total amount of insurance that customer purchased

'TotalDownPmt': Total amount of down payment

1. Our methodology:

* We have first combined datafiles. 1 to 3 into 1 file, 4 to 6 into another and lastly 7 to 9 into 1 file.
* We have joined above 3 files with a join condition on foreign keys for each.
* We calculated recovery rate from target table with a sql query. The logic was- average of least score indicator and second highest score indicator.
* We combined the final target table (consisting of recovery rate) with the combined data set files with other columns.
* We cleansed the data (removed duplicates, imputed the null values) and we identified appropriate columns for the model.

1. We created linear regression model with Xi as ('Vintage','ChargeOffMOB','BalanceAtDefault','FICOScore','FICOScorePctAvgFICOLast30DaysBookedLoans','NewUsedIndicator','VehicleManufacturerRebate','VehicleValueBlackBook','MoodysGDPReal','FinancedAmt','CashDownPmt','CashDownPmtPctBBVehicleValue','PaymentAmt','PaymentToFinancedAmtRatio','PTILine5','BackendTotalAmt','TotalDownPmt') columns and y as ‘recovery\_rate’.
2. We did a cross validation k=10 on linear regression model. We predicted the series of y values and our MAE is 16% , means the model mean absolute error is only 16%. Which means 84% of the time- it is going to predict Recovery rate correctly.