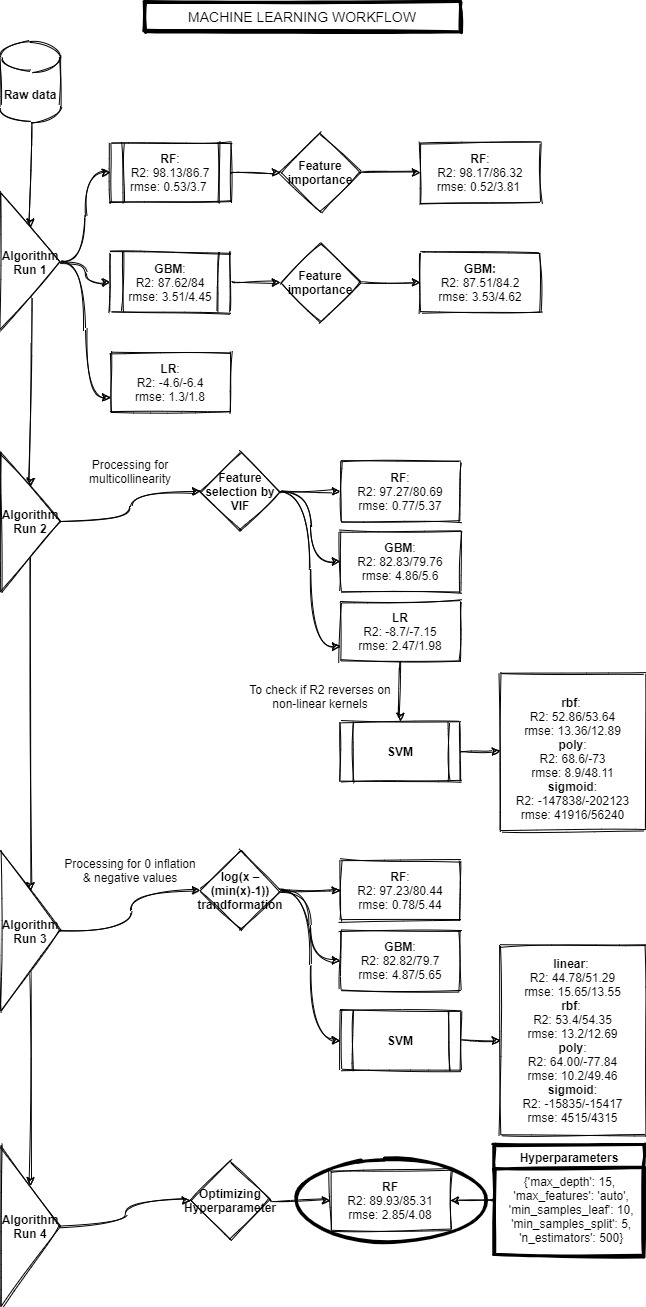
**The solution to predict the business risk of the customer**

****

**Step 1: Descriptive analysis of the train data with** [**Pandas Profiling**](profile_report_1__20210730-134015.html)

Following observations were noted based on the [Pandas Profiling](profile_report_1__20210730-134015.html)

|  |  |
| --- | --- |
| **Sr No** | **Observations** |
| 1 | There were no missing values |
| 2 | High multicollinearity |
| 3 | Zero-inflated data |
| 4 | Negative values present in several features |
| 5 | Several features were skewed |

**Step 2: Algorithm run 1: Application of brute force algorithms to know the efficiency of the models on raw data**

Following were the algorithms used with their respective R2 and RMSE values on train and val

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algorithm** | **R2** | | **RMSE** | | **Inference** |
| **Train** | **Val** | **Train** | **Val** |
| Random forest | 98.13 | 86.7 | 0.53 | 3.7 | Overfitting |
| Gradient boosting | 87.62 | 84 | 3.51 | 4.45 | Good Model |
| Stochastic gradient descent -linear regression | -4.6 | -6.4 | 1.3 | 1.8 | -R2, thus model does not fit well on the data |

**Step 3: Algorithm run 2: Random forest regressor and Gradient boosting regressor were run on respective top 10 important features to reduce dimensionality/complexity to check if there is an increase in the efficiency of the models on raw data**

Following were the algorithms used with their respective R2 and RMSE values on train and val

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algorithm** | **R2** | | **RMSE** | | **Inference** |
| **Train** | **Val** | **Train** | **Val** |
| Random forest | 98.17 | 86.32 | 0.52 | 3.81 | Overfitting |
| Gradient boosting | 87.51 | 84.2 | 3.53 | 4.62 | Good Model |
| Stochastic gradient descent -linear regression | -4.6 | -6.4 | 1.3 | 1.8 | -R2, thus model does not fit well on the data |

**Step 4: Treatment of multicollinearity based on variance inflation factor (VIF), threshold <5**

Following 41 columns were selected

|  |  |  |
| --- | --- | --- |
| '7days\_all\_gap\_days' | 'all\_7days\_min\_thisvs4w' | 'all\_norm\_growth\_m3' |
| 'all\_gap\_7days\_last\_vs\_previous' | 'all\_7days\_trend\_vs4weeks' | 'all\_norm\_growth\_m4' |
| '30days\_all\_gap\_days' | 'all\_7days\_trend\_vs10weeks' | 'all\_norm\_growth\_m5' |
| 'all\_gap\_30days\_last\_vs\_previous' | 'all\_7days\_vslast\_month7days' | 'all\_norm\_growth\_m6' |
| 'mtd\_all\_gap\_days' | 'all\_7days\_max\_thisvs10w' | 'all\_norm\_growth\_index\_last' |
| 'all\_gap\_mtd\_previous\_days' | 'all\_ystrday\_vsmin10d' | 'all\_gtv\_last12Months\_m12' |
| 'all\_last\_day' | 'all\_ystrday\_trend\_vs10d' | 'all\_gtv\_last10days\_d3' |
| 'all\_last30\_stable' | 'all\_ystrday\_vsdaybfr' | 'all\_gtv\_last10days\_d4' |
| 'all\_last30\_inc\_count' | 'all\_mrr\_trend\_vs6M' | 'all\_gtv\_last10days\_d6' |
| 'all\_consistency\_index' | 'all\_lst30days\_vsprvmnth' | 'all\_gtv\_last10days\_d7' |
| 'avg\_all\_gap\_days\_d1\_10' | 'all\_mtd\_vs\_min\_lst3M' | 'all\_gtv\_last10days\_d8' |
| 'avg\_all\_gap\_days\_d11\_20' | 'all\_trend\_mtdvs3M\_sameday' | 'all\_gtv\_last10days\_d9' |
| 'all\_gap\_days\_d1\_10\_thisvsprev' | 'all\_norm\_growth\_m1' | 'all\_norm\_growth\_m2' |
| 'all\_gap\_days\_d11\_20\_thisvsprev' | 'all\_gap\_days\_d20\_31\_thisvsprev' |  |

**Step 5: Algorithm run 3: Test for improvement in the model**

Following are the results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algorithm** | **R2** | | **RMSE** | | **Inference** |
| **Train** | **Val** | **Train** | **Val** |
| randomForest regressor | 97.27 | 80.69 | 0.77 | 5.37 | Overfitting |
| Gradient Boosting regressor | 82.83 | 79.76 | 4.86 | 5.6 | Good Model but performance slightly down from the brute force |
| SVR (rbf) | 52.86 | 53.64 | 13.36 | 12.89 | Highly robust since no variation in train and val, however not as good as GBM |
| SVR (poly) | 68.6 | -73 | 8.9 | 48.11 | Poor model |
| SVR (sigmoid) | -147838.86 | -202123 | 41916.05 | 56240.87 | A poor model since R2 is negative |
| SVR (linear) | 45.32 | 54.38 | 15.49 | 12.69 | + R2 with normalized data, however, the model is not as good as GBM |

**Step 6: Algorithm run 4: Data was normalized with log(x – (min(x)-1)) since the data had negative values as well as zero inflation.**

Following are the results for VIF + normalized data.

Further, since R2 was negative for linear regression, support vector machine methods with different kernels were tested.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algorithm** | **R2** | | **RMSE** | | **Inference** |
| **Train** | **Val** | **Train** | **Val** |
| randomForest regressor | 97.23 | 80.44 | 0.78 | 5.44 | Overfitting |
| Gradient Boosting regressor | 82.82 | 79.7 | 4.87 | 5.65 | No improvement with the normalized data |
| SVR (rbf) | 53.4 | 54.35 | 13.2 | 12.69 | Highly robust since no variation in train and val, however not as good as GBM |
| SVR (poly) | 64 | -77.84 | 10.2 | 49.46 | Poor model |
| SVR (sigmoid) | -15835.59 | -15417.5 | 4515.09 | 4315.64 | A poor model since R2 is negative |
| SVR (linear) | 44.78 | 51.29 | 15.65 | 13.55 | + R2 with normalized data, however, the model is not as good as GBM |

**Step 7: Hyperparameter tuning of random forest and gradient boosting regressor**

Looking at all the previous readings, random forest and gbm with respective top 10 important features gave better results wrt better generalization (gbm) and error rate (gbm + rf). Hence, the fine-tuning of these algorithms for the hyperparameters were to be carried out, Hyperparameter tuning for random forest improved the overfitting issue, however, due to time constraints and computation limitations, the same for gradient boosting regressor could not be concluded at this stage.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algorithm** | **R2** | | **RMSE** | | **Inference** |
| **Train** | **Val** | **Train** | **Val** |
| Random forest | 89.93 | 85.31 | 2.85 | 4.08 | Improvement in the model due to hyperparameter tuning. The overfitting issue has reduced significantly without much affecting the performance |

**Best params:**

**'max\_depth'**: 15, **'max\_features'**: 'auto', **'min\_samples\_leaf'**: 10, **'min\_samples\_split'**: 5, **'n\_estimators'**: 500

**Step 8: Feature importance of random forest regressor as follows:**

|  |  |
| --- | --- |
| **Feature - RF** | **Importance** |
| 3M\_daily\_all\_avg | 0.193174 |
| 3M\_weekly\_all\_avg | 0.187331 |
| all\_lst30days\_vsprvmnth | 0.183720 |
| all\_last\_day | 0.082440 |
| all\_lst30days\_vsmax\_lst3m | 0.075532 |
| all\_lst30days\_vsmean\_lst3m | 0.049418 |
| all\_7days\_trend\_vs4weeks | 0.044291 |
| all\_gtv\_last10weeks\_w2 | 0.017866 |
| all\_gtv\_last12Months\_m1 | 0.013820 |
| all\_7days\_max\_thisvs10w | 0.013073 |

**Conclusion:**

Risk management is the process to identify threats to an organization’s revenue cycle. Considering this principle, in the given business usecase, as the three month’s daily or weekly sales are up and even last 30 days sales against the previous month are up, the more is the threat to the company if that customer leaves the company. Thus such predictions will help the company to prepare for the unexpected event by assessing such customers.