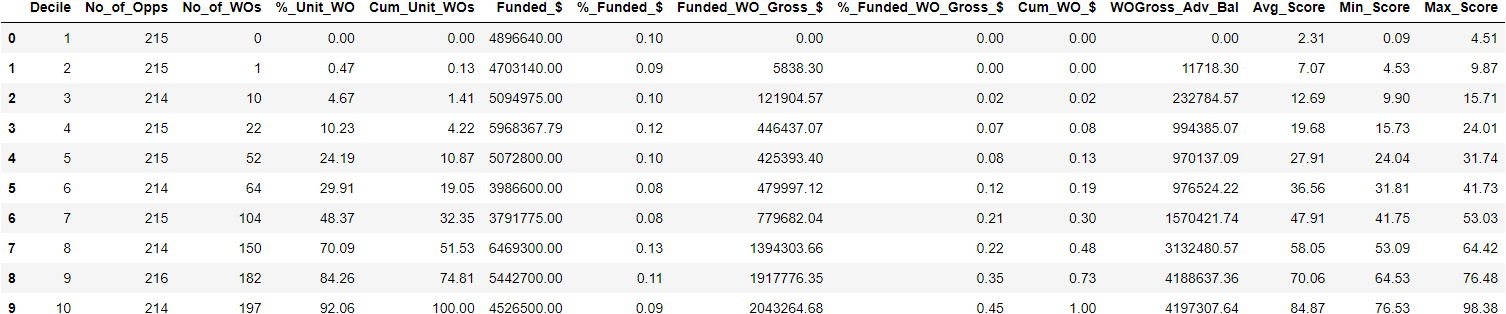
| **Data Categories** | **2019** | **2020** | **2021** | **2022**  **(Jan. 31st)** | **Total** | **Overall %Unit\_WO\_Rate** |
| --- | --- | --- | --- | --- | --- | --- |
| **Ren + LexisNexis + CBC**  **(Original Data)** | (269, 666) | (765, 666) | (1083, 666) | (78, 666) | (2195, 666) | 22% - (489)  78% - (1706) |
| **Ren + LexisNexis + CBC**  **(Upsampling\_1)** | (323,666) | (902,666) | (1358,666) | (101,666) | (2684,666) | 36% - (978)  64% - (1706)  Training - (2147,656)  Val - (537,656) |
| **Ren + LexisNexis + CBC**  **(Upsampling\_2)** | (377,666) | (1039,666) | (1633,666) | (124,666) | (3173,666) | 46% - (1467)  54% - (1706)  Training - (2538,656)  Val - (635,656) |

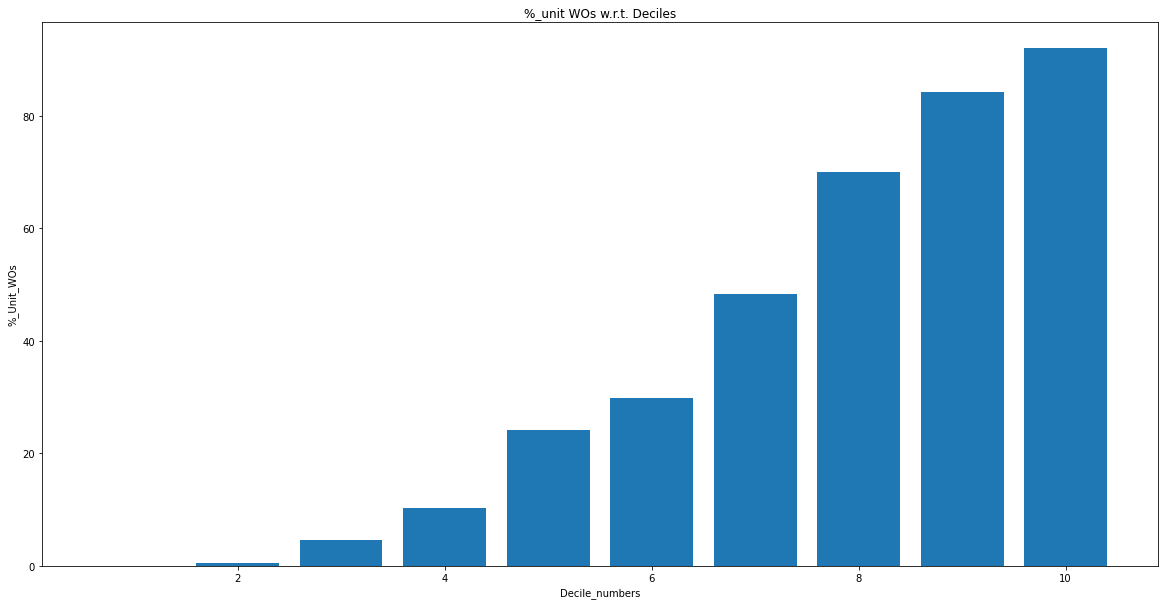
**Model 1 After upsampling\_1**

* **We doubled the positive classes (WriteOff\_YN = 1).**
* **Earlier the class balance was 22:78. Now the ratio is 36:64.**
* **Data shape - (2684, 666)**
* **Shape of x\_train is: (2147, 656)**
* **Shape of x\_val is: (537, 656)**
* **Data\_2019 - (323, 666)**
* **Data\_2020 - (902, 666)**
* **Data\_2021 - (1358, 666)**
* **Data\_2022 - (101, 666) (Upto 31st January)**

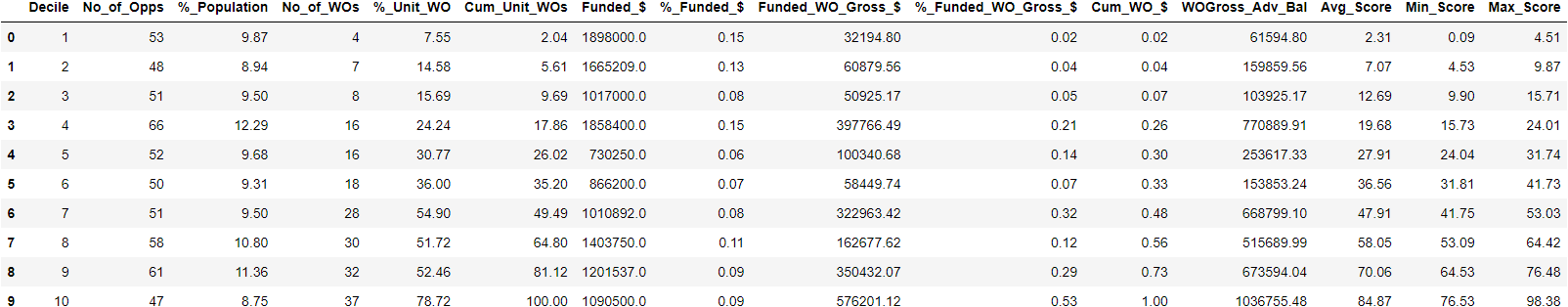


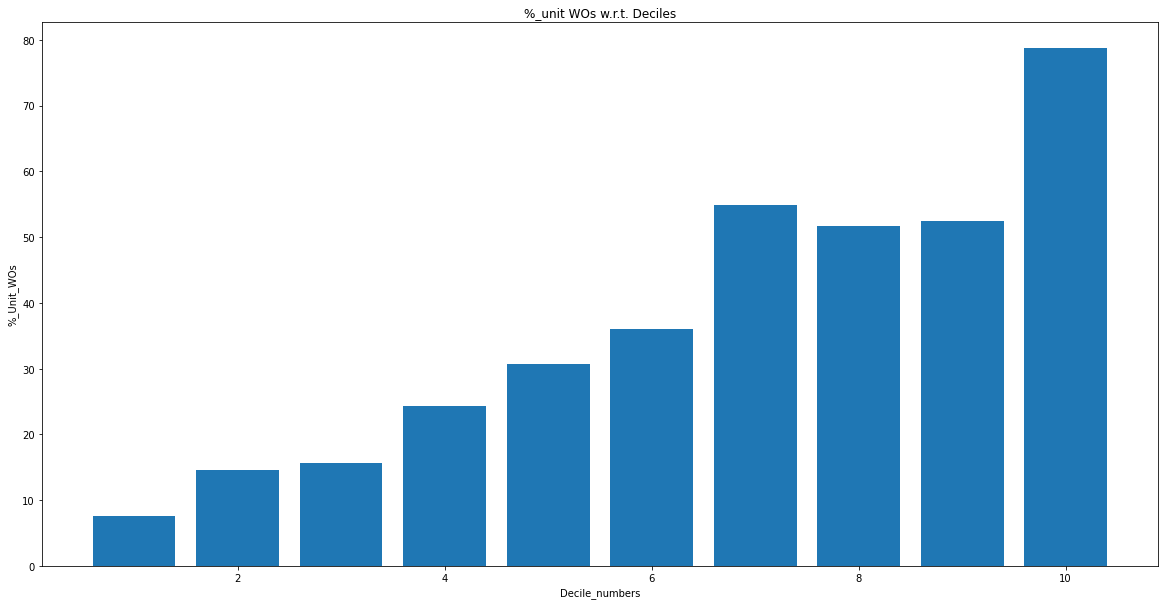
Train set Deciles



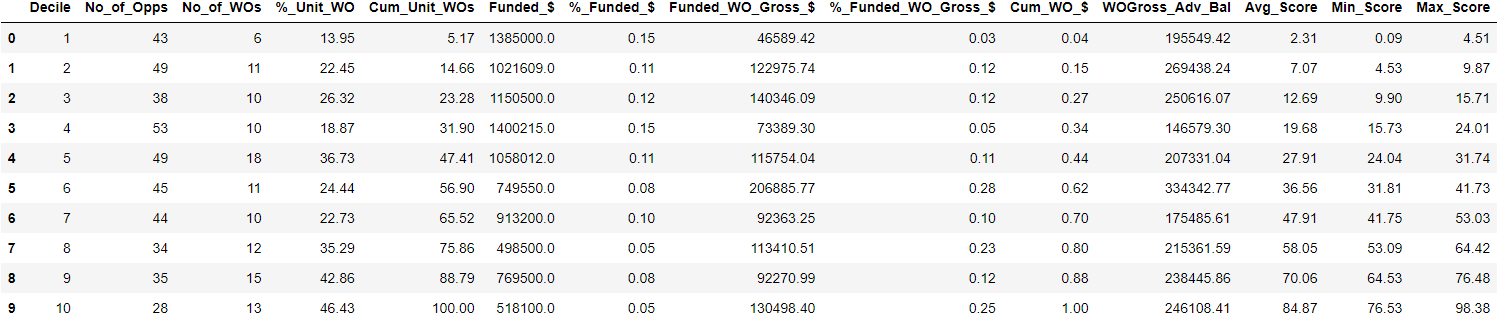


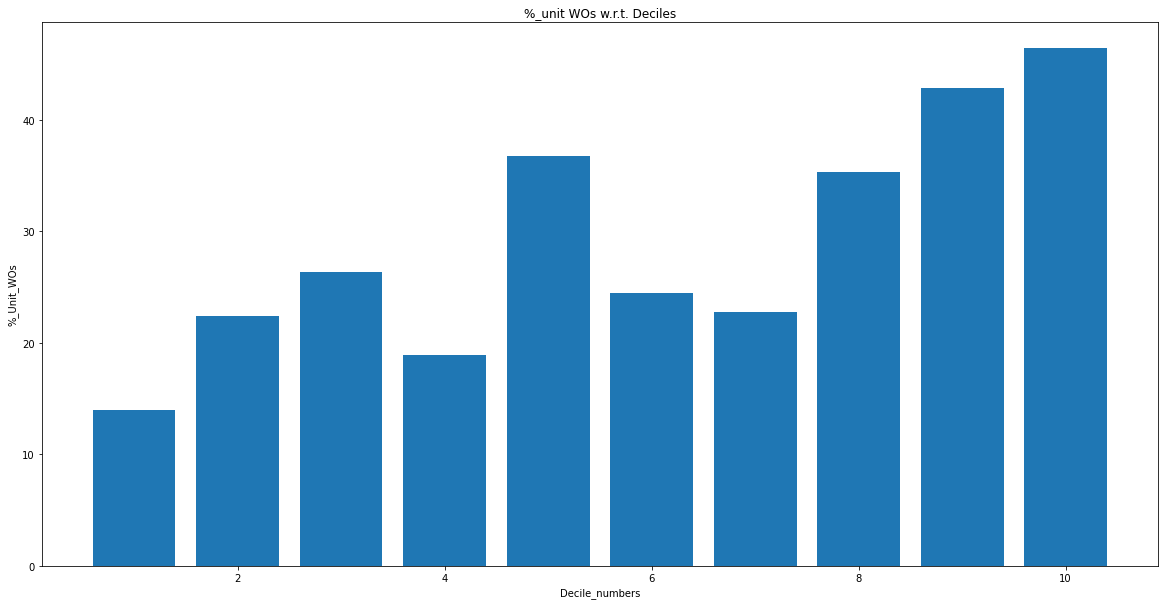
Val Set Deciles



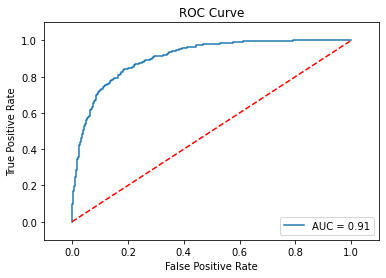


Test set Deciles

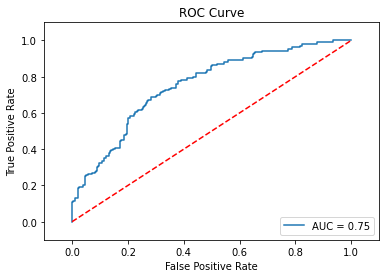




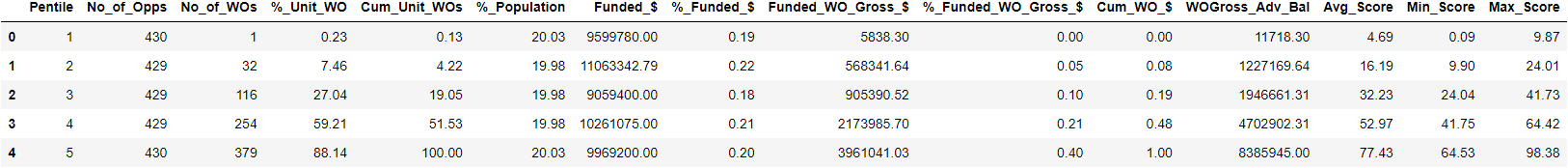
Training AUC\_ROC

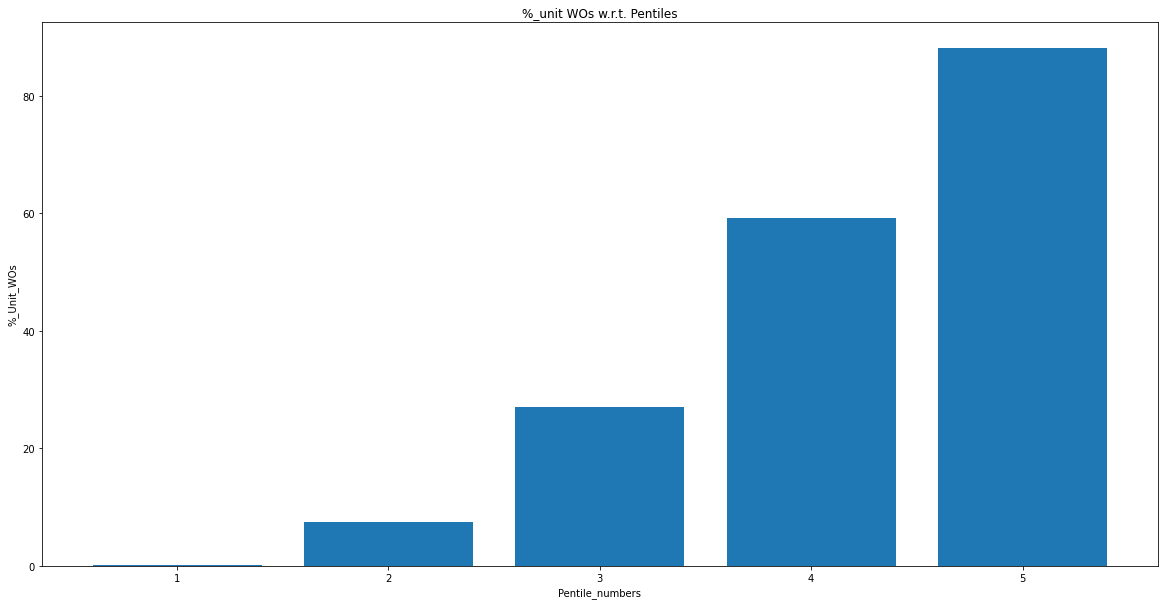


Val AUC\_ROC

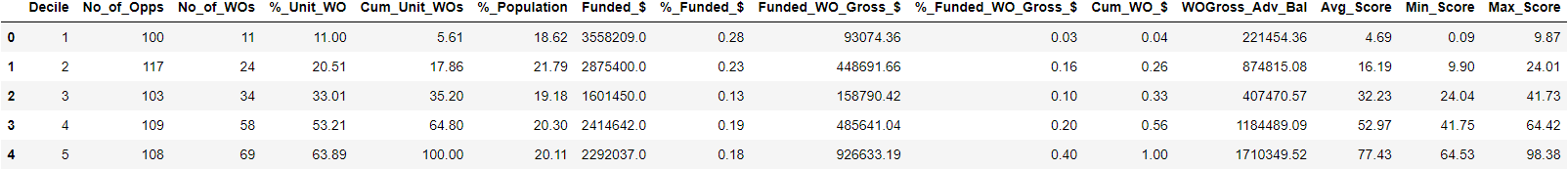


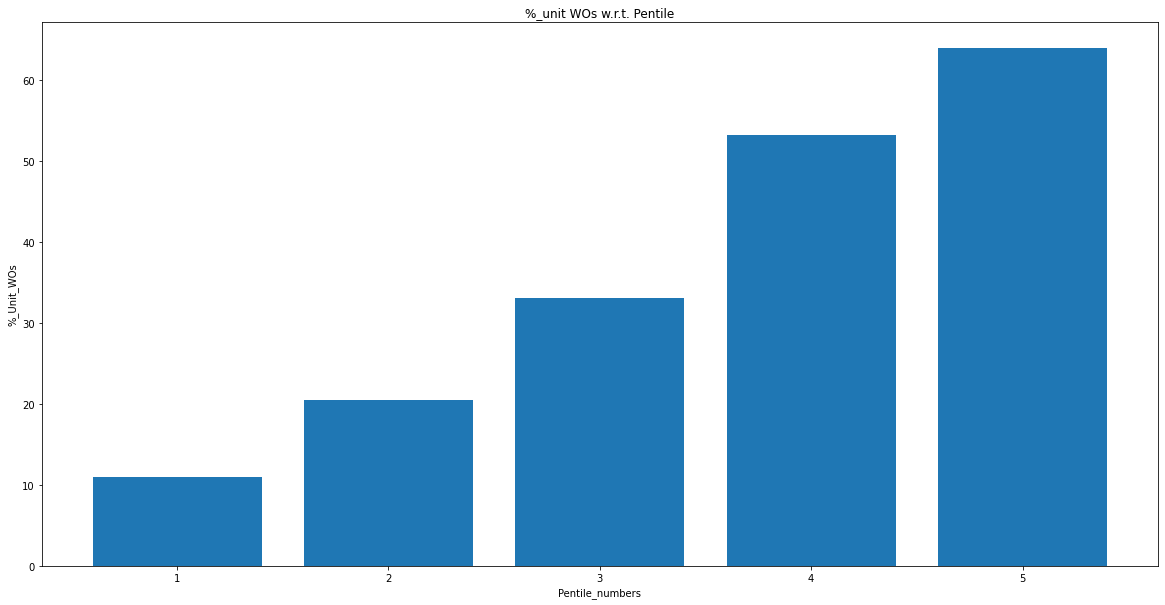
Train Set Pentile



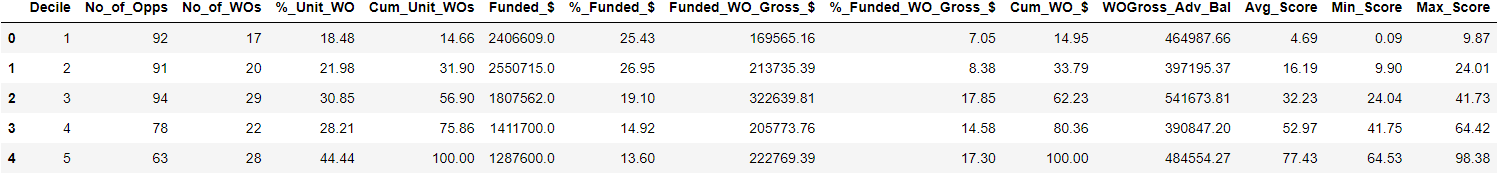


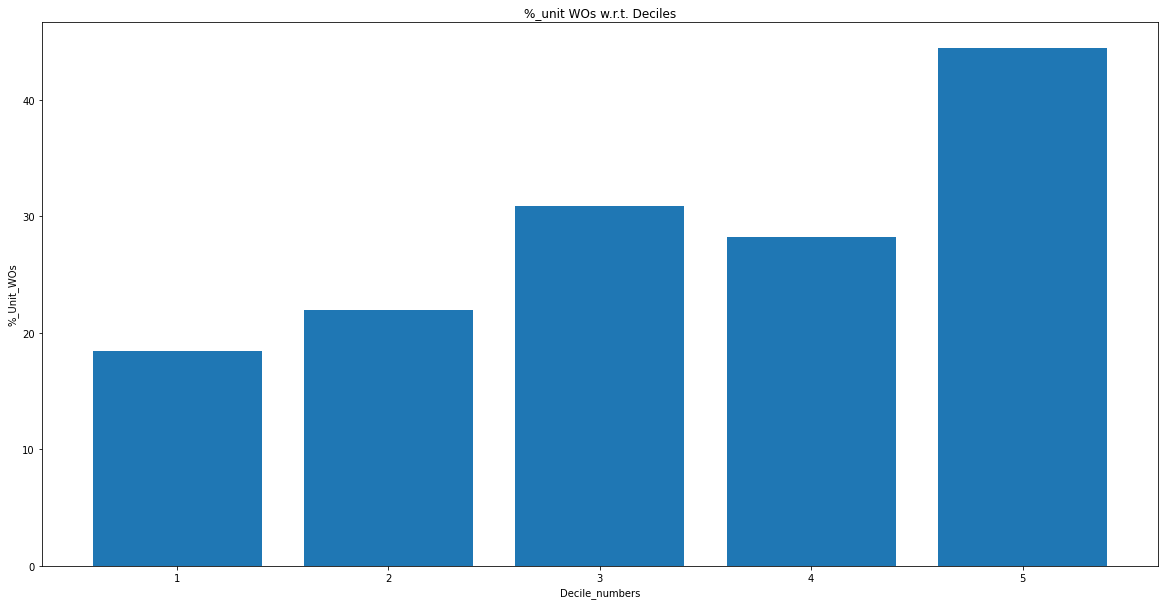
Val Set Pentile





Test set Pentile



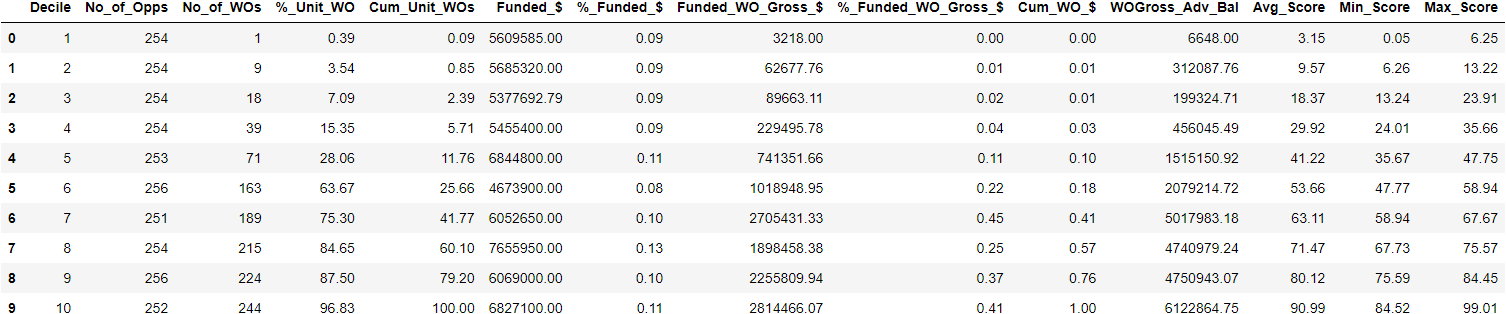


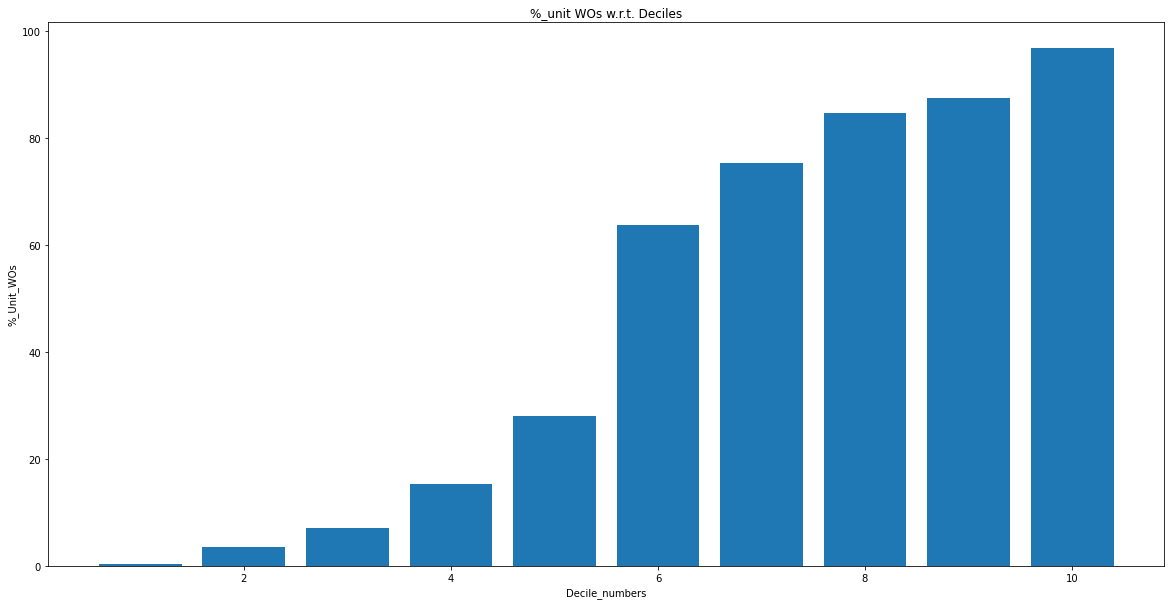
**Model 2 After upsampling\_2**

* **We tripled the positive classes (WriteOff\_YN = 1).**
* **Earlier the class balance was 22:78. Now the ratio is 46:54.**
* **Data shape - (3173, 666)**
* **Shape of x\_train is: (2538, 656)**
* **Shape of x\_val is: (635, 656)**
* **Data\_2019 - (377, 666)**
* **Data\_2020 - (1039, 666)**
* **Data\_2021 - (1633, 666)**
* **Data\_2022 - (124, 666) (Upto 31st January)**

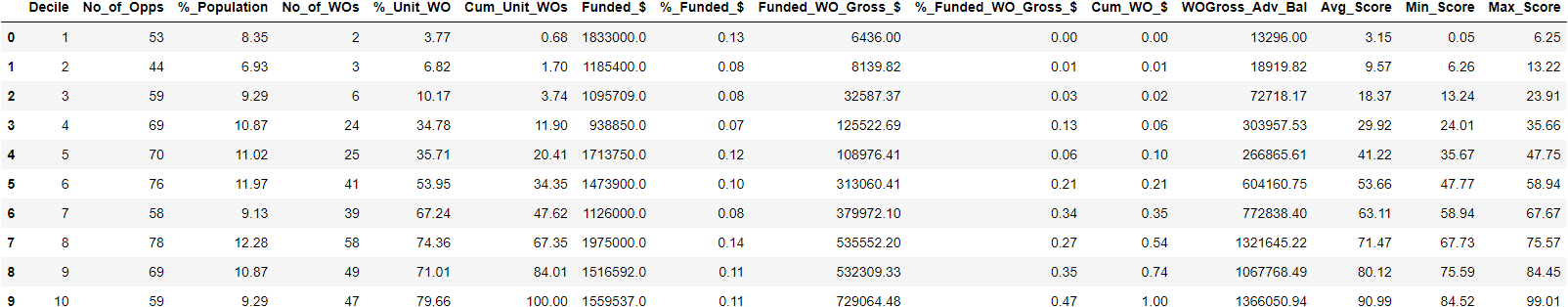


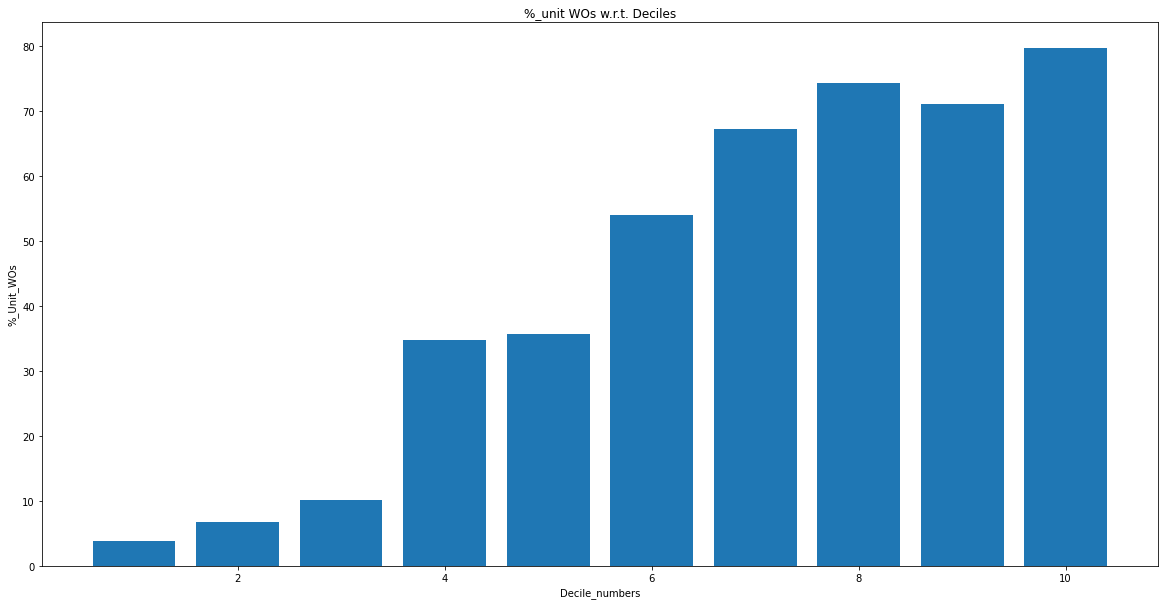
Train set Decile



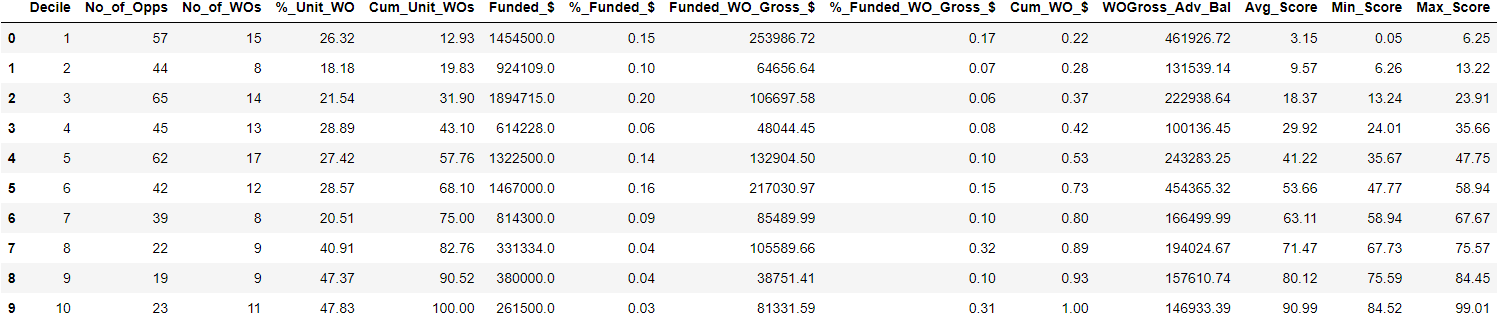


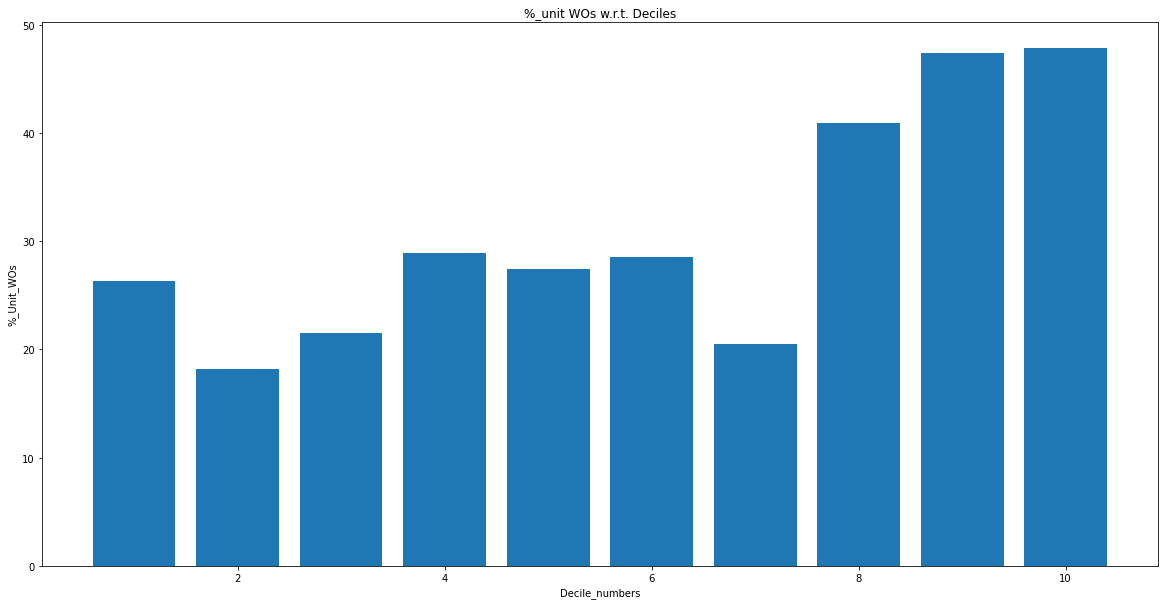
Val set Decile



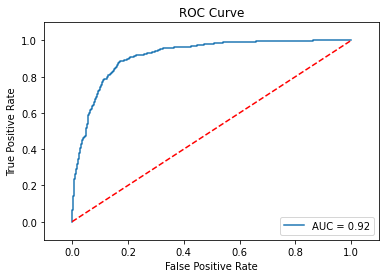


Test Set Decile

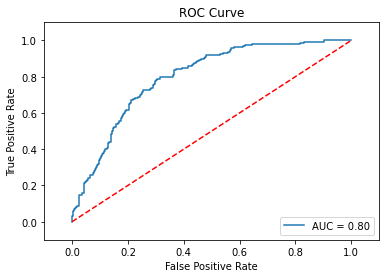




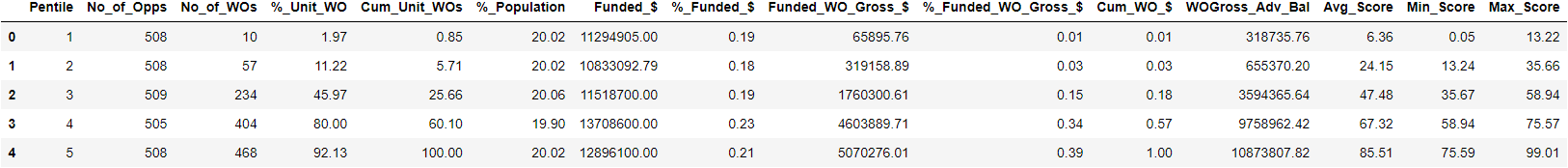
Train set AUC\_ROC

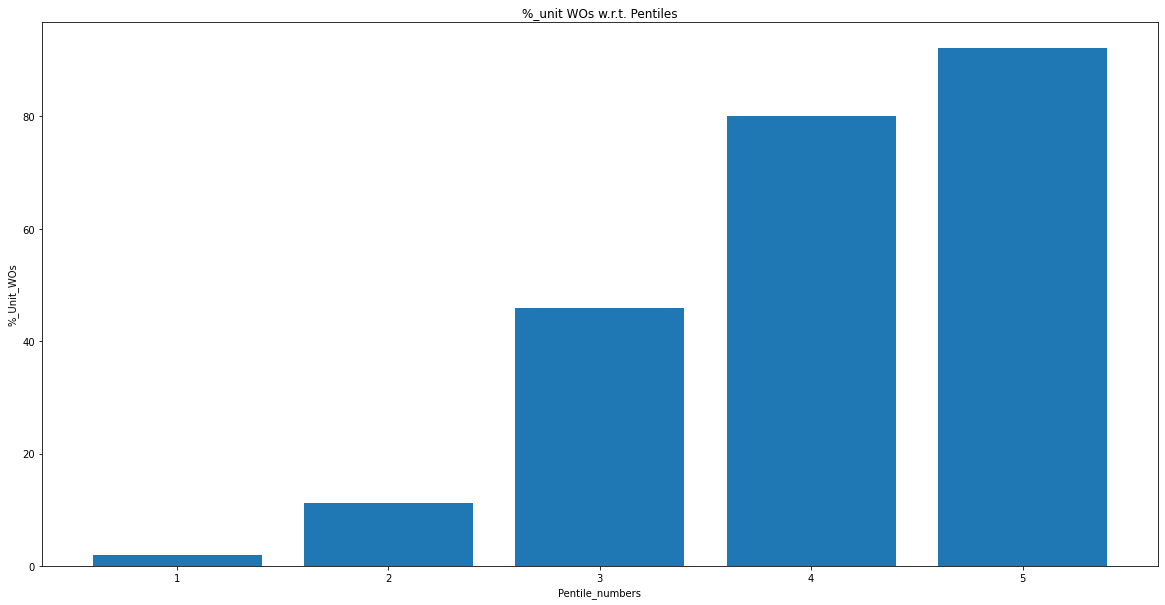


Val set AUC\_ROC

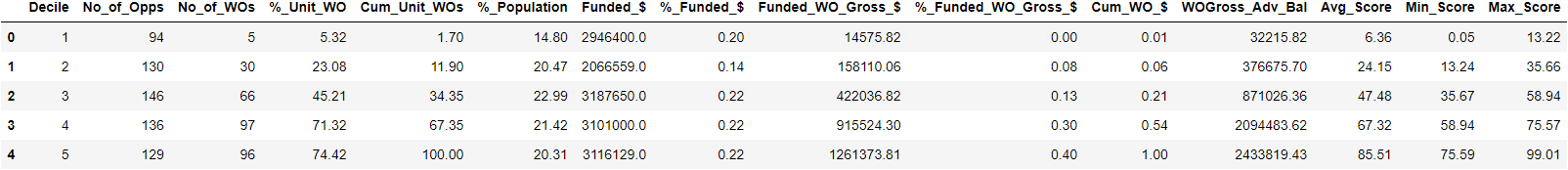


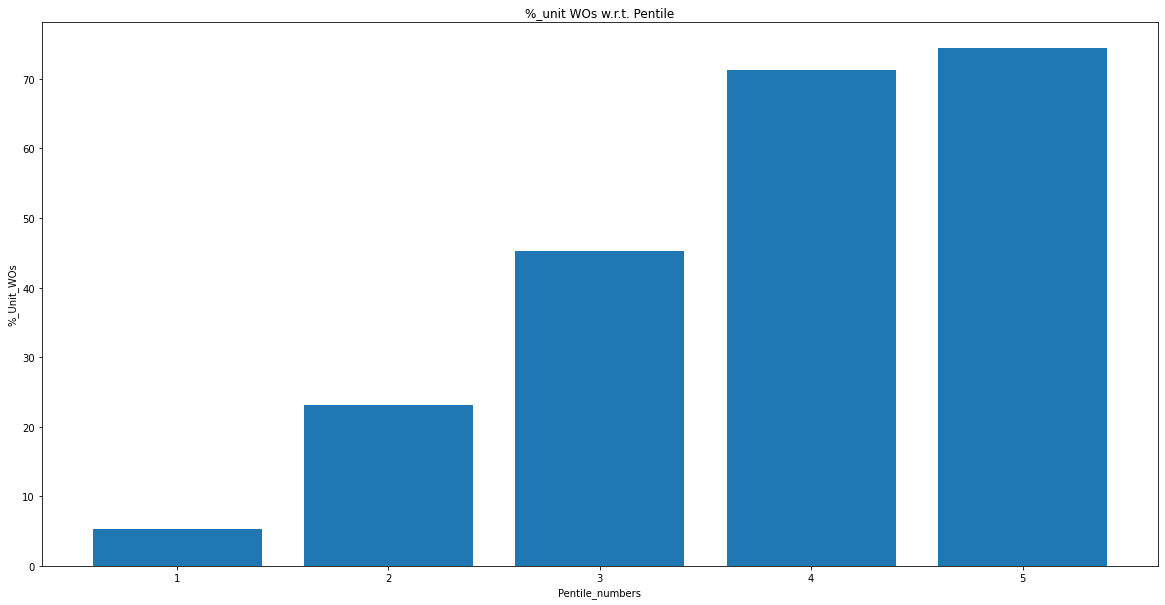
Train set Pentile



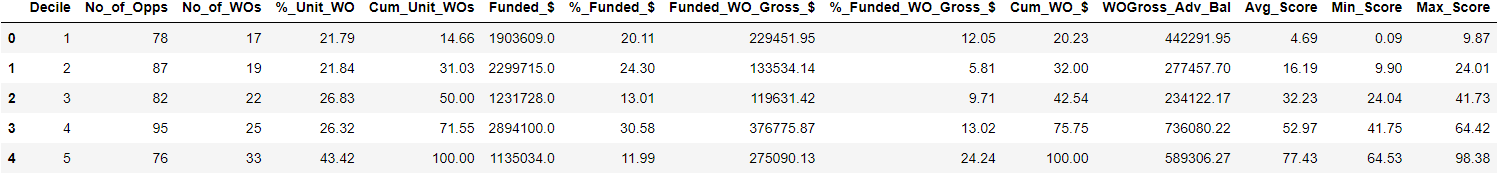


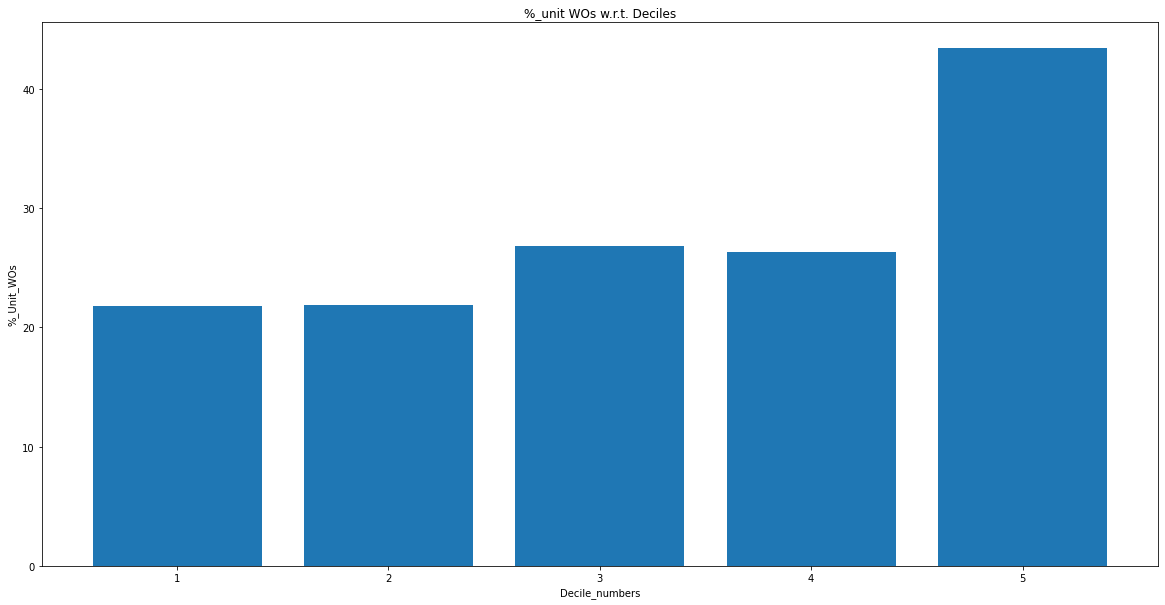
Val set Pentile





Test set Pentile

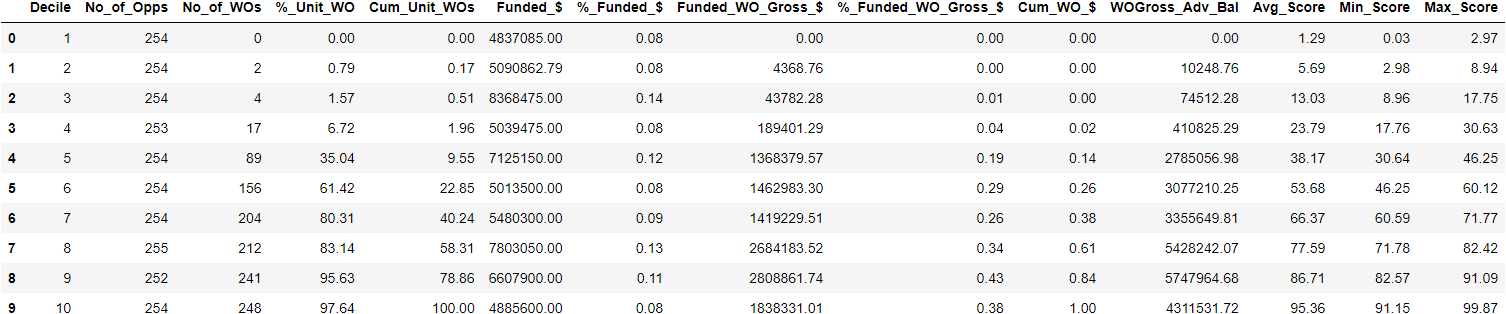


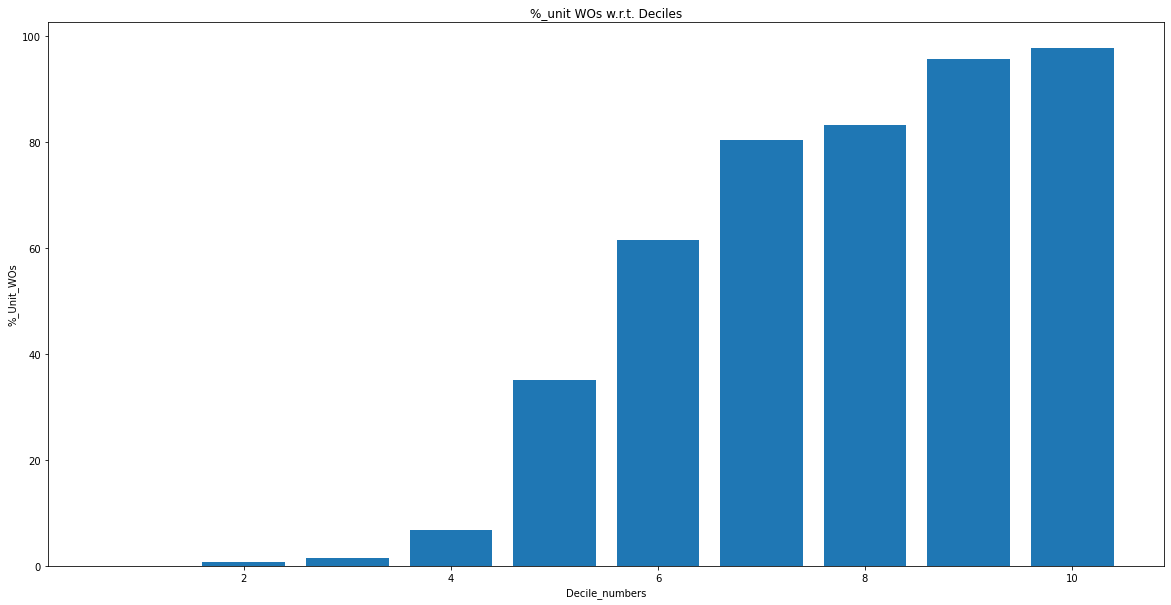


Model 3 after upsampling\_2 (same data distribution as previous)

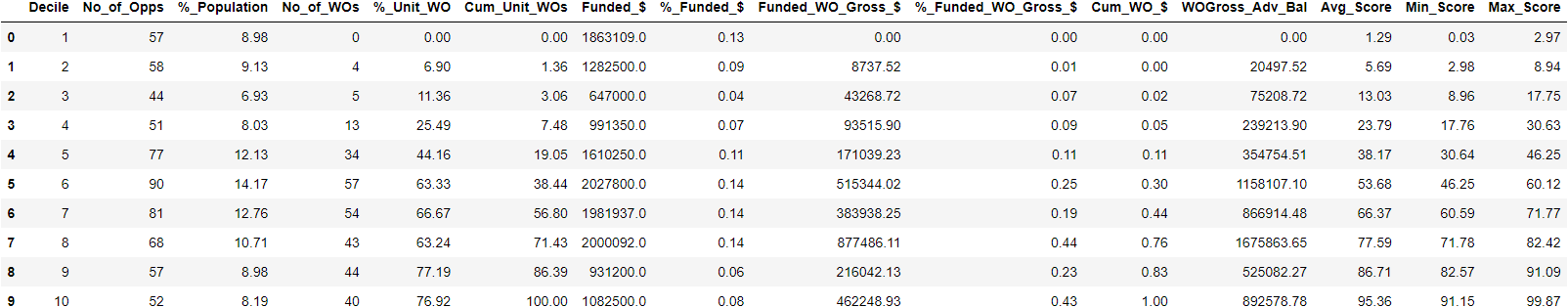


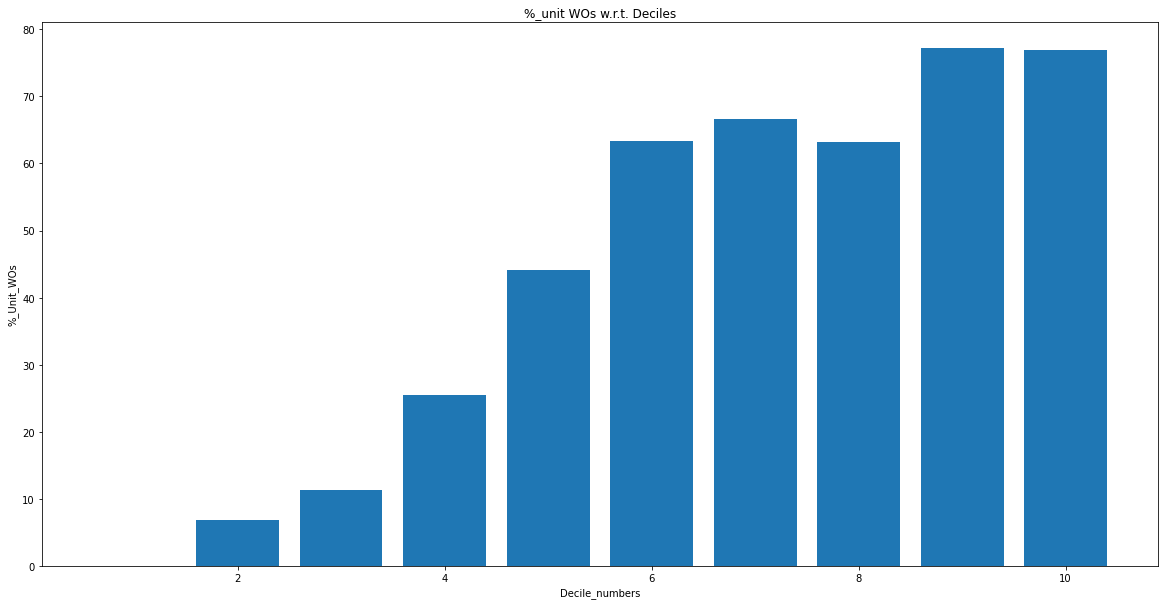
Train set Decile



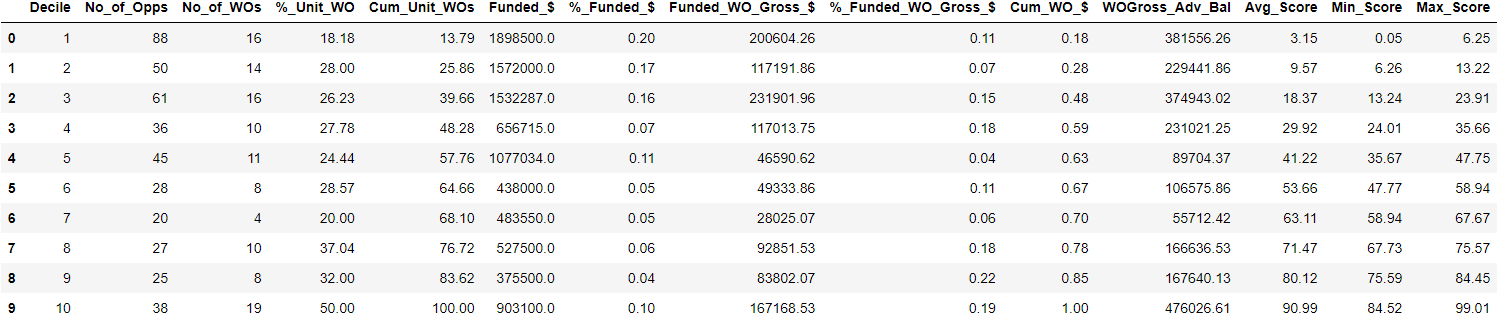


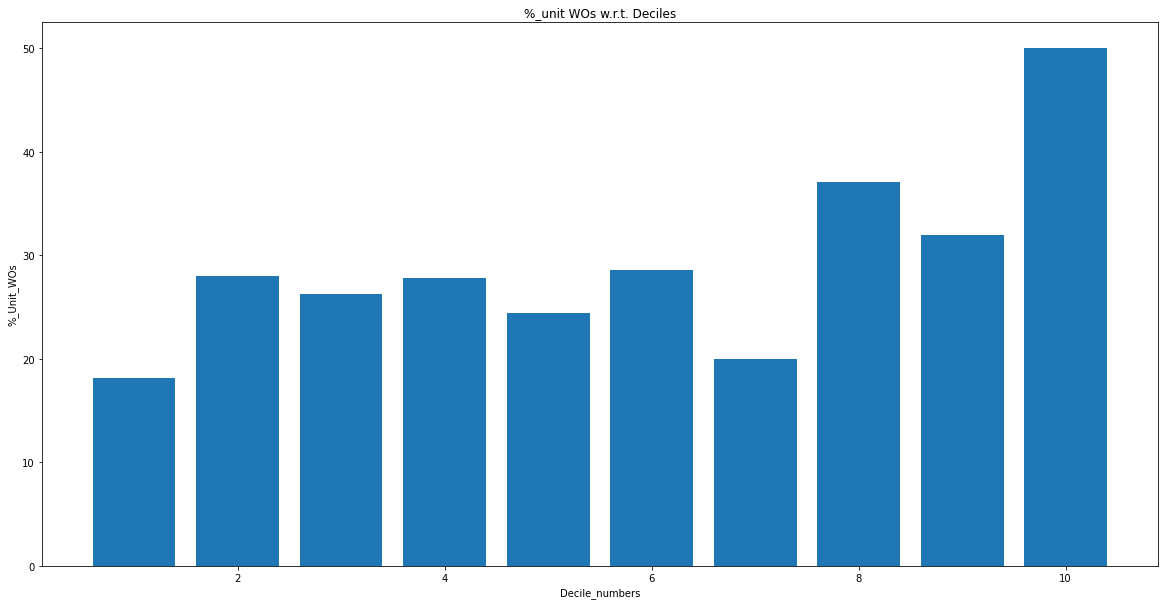
Val set Decile



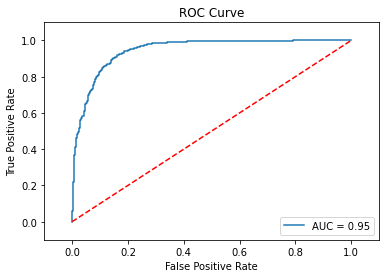


Test set Decile

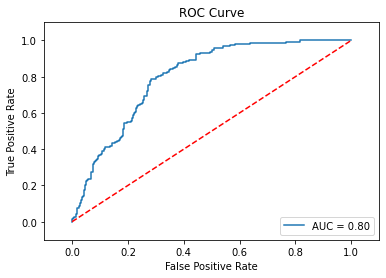




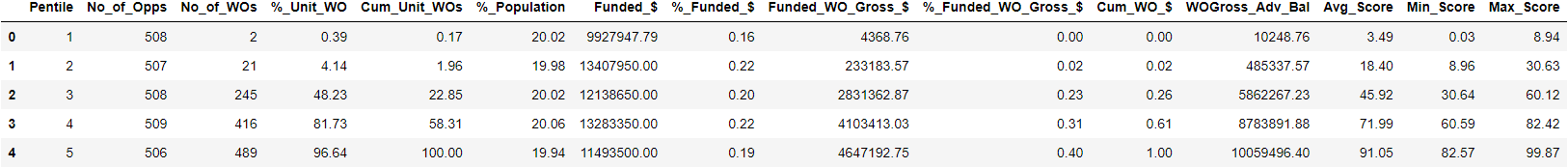
Train AUC\_ROC

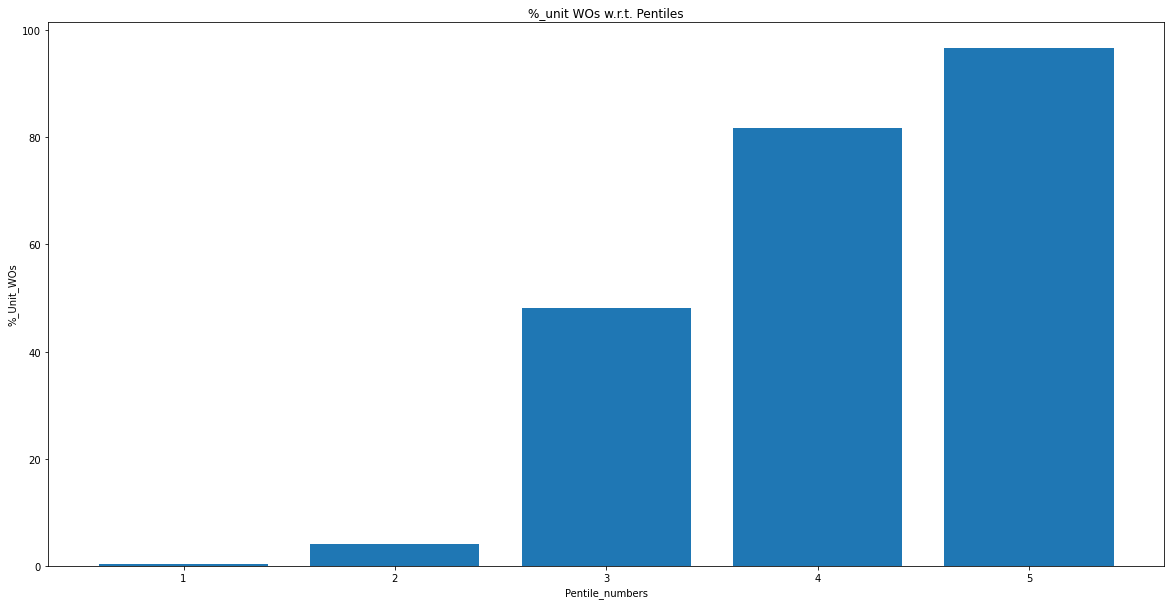


Val AUC\_ROC

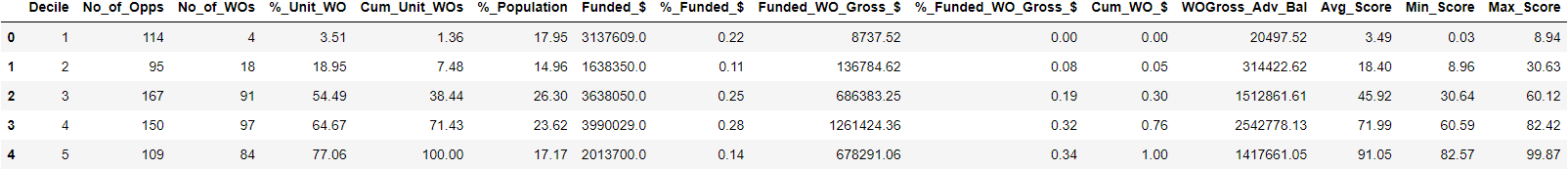


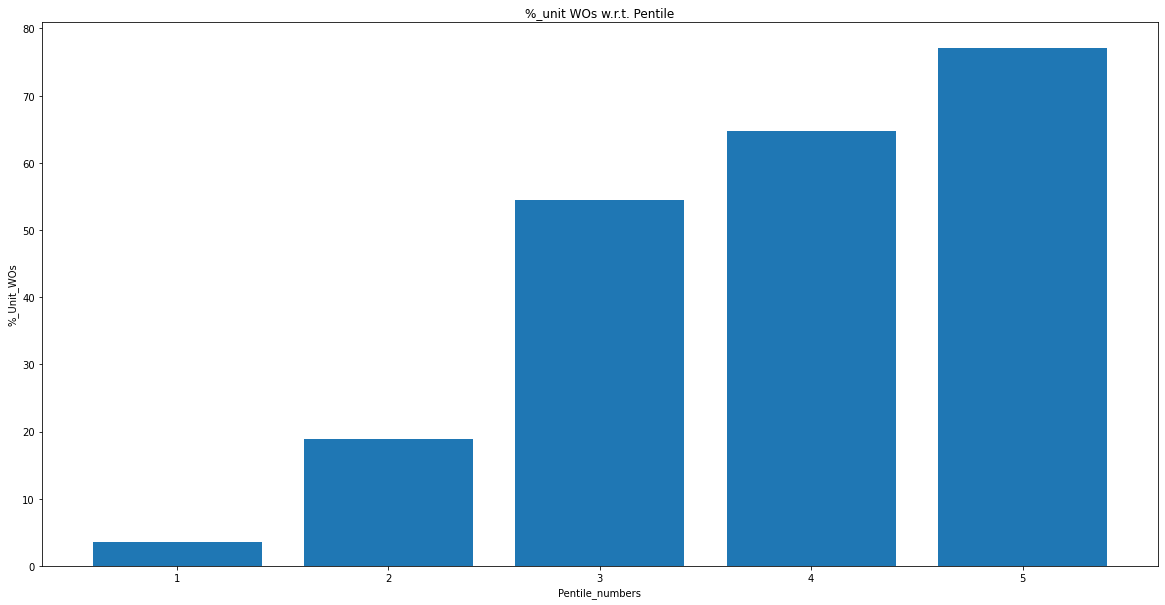
Train set Pentile

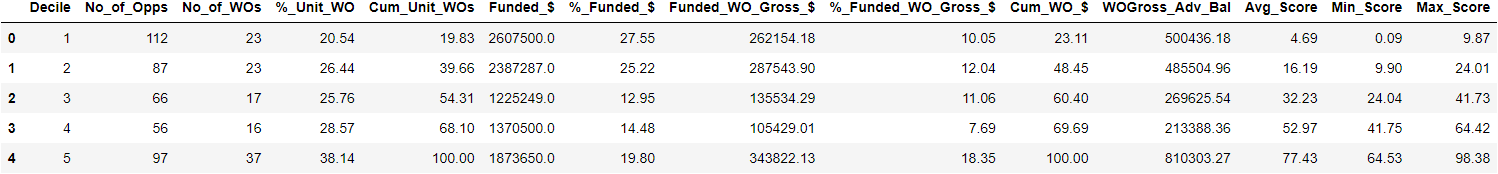




Val set Pentile





Test set Pentile

**What we have covered so far:**

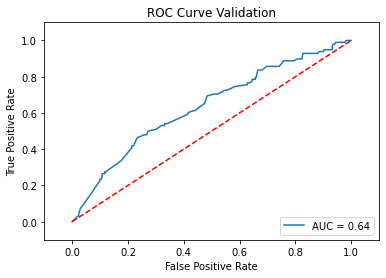
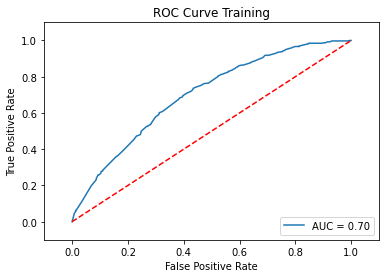
* EDA on REN, LEN, CBC Data
  + We removed unwanted data-points
  + Dropped categorical attributes
* For feature engineering
  + Extra Features Created
  + Feature Scaling
* Model Training Categories
  + REN
  + REN\_LN
  + REN\_CBC\_LEN
* Model Training Experiments
  + First on all attributes
  + Taking top 50 attributes and then training model
  + Dropping attributes based on correlations followed by model training
  + Giving special weightage to WO\_Gross\_Advance\_Balance followed by training
  + Giving Special Weightage to False Negativesin upper deciles followed by training
  + Upsampling to balance positive and negative output classes
* Hyper-parameter Tuning
  + Using Cross Validation
  + Controlling output class imbalance using scale pose weight
  + Giving special weightage to WO\_Gross\_Advance\_Balance followed by training
  + Giving Special Weightage to False Positive followed by training
  + L1, L2, Learning\_Rate, Depth, N\_Estimators are tweaked followed by decile analysis
* Basic Model Training on Logistic Regression
  + Hyper-parameter Tuning
  + Decile creation

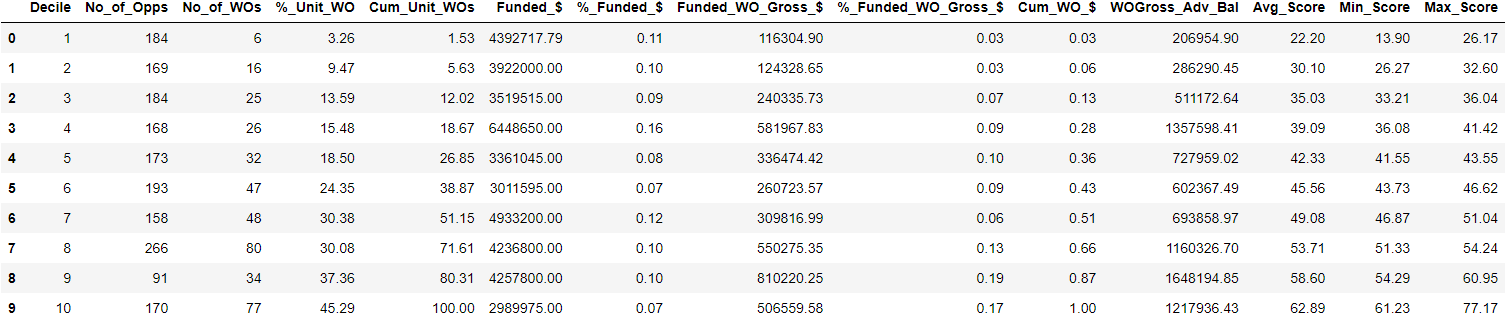
Further Experiments:

* Extensive training on Logistic Regression
* **T-SNE**
  + For visualizing high-dimensional data in two dimensions.
  + Very useful for error analysis.
  + E.g. class separation (give each class a separate color)
  + inter-decile separation (give each decile a separate color)
  + intra-decile separation
  + <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html>
  + <https://builtin.com/data-science/tsne-python>
* **Model Selection and Evaluation** 
  + Framework for model selection and evaluation:
  + <https://scikit-learn.org/stable/model_selection.html>
  + User-defined evaluation metric:
  + <https://scikit-learn.org/stable/modules/model_evaluation.html>
  + See Section 3.3.1.3. Implementing your own scoring object.
  + By defining the task-specific evaluation metric and implementing it as a scoring object, you can automate model selection and evaluation completely. The task-specific evaluation metric can be a composite of classification metrics like F1 score and task-specific KPIs like cumulative gross writeoff.
* **Calibration (Calliberated\_CV)**
  + Useful sometimes for getting better estimates of prediction probabilities.
  + <https://scikit-learn.org/stable/modules/calibration.html>
  + <https://www.youtube.com/watch?v=AunotauS5yI>
  + ​​<https://youtu.be/5zbV24vyO44>
* **Logistic Regression**
  + <https://regressit.com/regressitlogistic.html>
  + Very useful tool for doing LR analysis in Excel with useful visualization and statistics.
  + Also useful for catching errors in the data early.
  + Also useful for p value based feature selection.

Calibrated\_CV approach

Model Attribute category - REN\_CBC\_LN

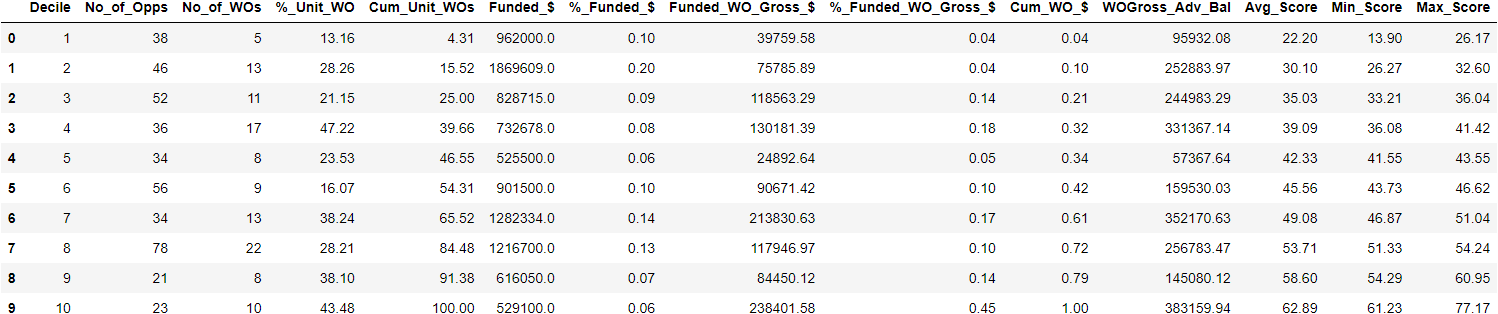
Check for overfitting condition (Train and Validation Data)

Training Decile without Caliberated\_CV

Training Decile with calibrated\_cv

Validation Decile without caliberated\_cv

Validation Decile with calibrated\_cv

Test Decile without calibrated\_cv

Test Decile with calliberated\_cv 