Error Generating Reflection on Data Modeling

Research Experiments

1) Random Domain Active Error Generating on the data (Anime Type)
2) Typo Error Generating in the data (Anime Genre)
3) Gaussian Noise in the highest ranked feature (Members who voted to the Anime)

RAD - Anime Type

1) **Generating**

   Shuffling the Anime Type (Tv series, Movie, NaN)

2) **Detecting**

   Depending on the Episodes numbers

   one Episodes -> Movie, otherwise it is Tv-series

3) **Repairing**

   Using Episodes numbers

RAD - Error on 25% of the Training Dataset

	Dirty Data		Cleaned Data	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	1.08	-18.18	0.86	0.10
Lasso With HP	0.84	-18	0.82	0.09

RAD - Error on 50% of the Training Dataset

	Dirty Data		Cleaned Data	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	1.12	-21.68	0.85	0.12
Lasso With HP	0.85	-18.56	0.85	0.10

RAD - Error on 75% of the Training Dataset

	Dirty Data		Cleaned Data	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	1.13	-20.17	0.85	0.15
Lasso With HP	0.87	-17	0.84	0.10

Typo - Anime Genre

1) **Generating**

   Create a Typo error in each tuple of data gener (eg. romantic → ronctzs)

2) **Detecting**

   Using spelling checker python library to detect the misspelled words

3) **Repairing**

   Correct the misspelled words, However, there are some strange words that could not correct them.

Typo - Error on 25% of the Training Dataset

	Dirty Data		Cleaned Data	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	1.10	-16.62	0.85	0.15
Lasso With HP	0.84	-12.56	0.85	0.15

Typo - Error on 50% of the Training Dataset

	Dirty Data		Cleaned Data	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	1.11	-30	0.88	0.10
Lasso With HP	0.82	-23.56	0.81	0.05

Typo - Error on 75% of the Training Dataset

	Dirty Data		Cleaned Data	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	1.14	-45.77	0.86	0.09
Lasso With HP	0.82	-19.55	0.81	0.05

Gaussian Noise - Voting Members

1) **Generating**

   Apply Gaussian Noise (u=0, sigma=10000) on the members to check the effects of Gaussian signal on the distribution of the most correlated feature.

2) **Detecting**

   Plotting the distribution shows us there is a noise in the data

3) **Repairing**

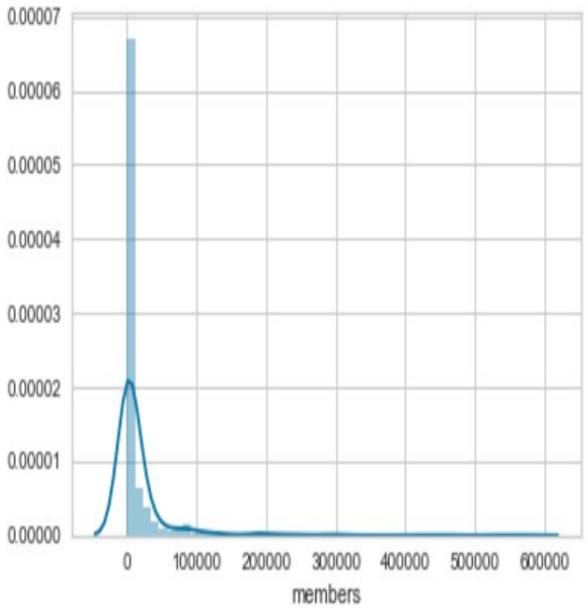
   Using savgol_filter to filter the data from the noise

GN - Error on 30% of the Training Dataset

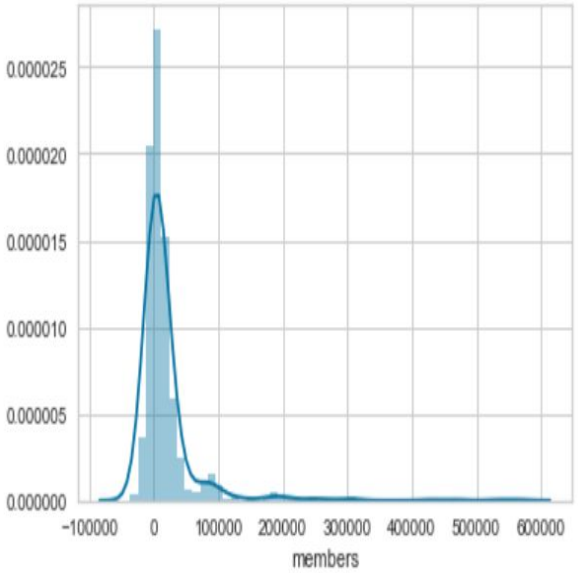
	Before		After	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	0.64	-1.34	0.97	-0.01
Lasso With HP	0.04	-0.03	0.94	0

Members Attribute Before and After Filtering - 30%

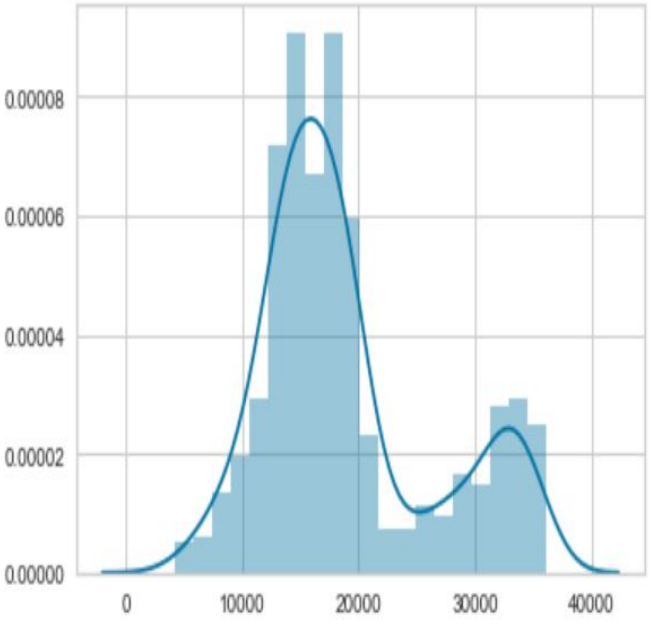
B-Gaussian Noise



W-Gaussian Noise



Filter-Gaussian Noise

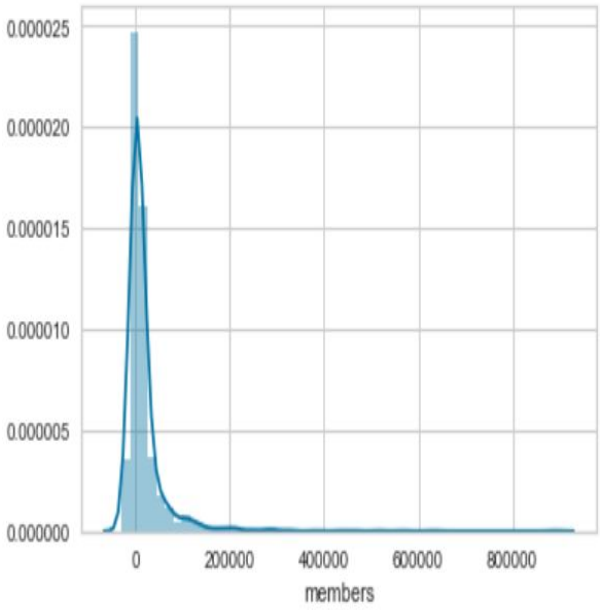


GN - Error on 100% of the Training Dataset

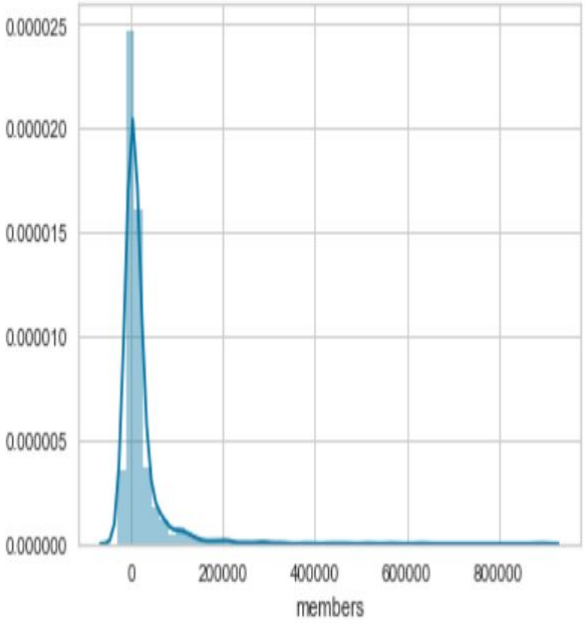
	Before		After	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	0.68	-3.96	0.97	-0.01
Lasso With HP	0.05	-0.02	1.86	-0.06

GN - Members Attribute Before and After Filtering - 100%

B-Gaussian Noise



W-Gaussian Noise



Filter-Gaussian Noise