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	0.85	0.08	0.83	0.18
Lasso With HP	0.83	0.16	0.83	0.18

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	0.90	-0.01	0.90	-0.01
Lasso With HP	0.89	0.02	0.87	0.14

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	0.87	0	0.87	0
Lasso With HP	0.82	0.18	0.83	0.11

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.

Generate a dictionary contains a list of most famous movies genre and check the similarity

Typo - Error in 25% - build-in corrector

	Dirty Data		Cleaned Data	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	1.28	-31.06	0.89	0
Lasso With HP	1.08	-17.16	0.85	0.15

Typo - Error in 50% - build-in corrector

	Dirty Data		Cleaned Data	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	1.30	-46.75	0.87	0
Lasso With HP	1.13	-30.38	0.84	0.12

Typo - Error in 75% - build-in corrector

	Dirty Data		Cleaned Data	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	1.29	-64.78	0.89	0
Lasso With HP	1.14	-42.36	0.86	0.14

Typo - Error in 25% - Dict with most common genre

	Dirty Data		Cleaned Data	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	1.28	-31.06	0.89	0
Lasso With HP	1.01	-17.16	0.84	0.15

Typo - Error in 50% - Dict with most common genre

	Dirty Data		Cleaned Data	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	1.30	-46.75	0.91	0
Lasso With HP	1.05	-20.09	0.87	0.14

Typo - Error in 75% - Dict with most common genre

	Dirty Data		Cleaned Data	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	1.32	-64.78	0.89	0
Lasso With HP	1.14	-42.36	0.86	0.14

Typo - Error in 25% - Drop Uncorrected Words

	Dirty Data		Cleaned Data	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	1.28	-60	0.91	0
Lasso With HP	1.12	-40	0.87	0.14

Typo - Error in 50% - Drop Uncorrected Words

	Dirty Data		Cleaned Data	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	1.27	-35	0.91	0
Lasso With HP	1.11	-30	0.86	0.16

Typo - Error in 75% - Drop Uncorrected Words

	Dirty Data		Cleaned Data	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	1.32	-64.78	0.89	0
Lasso With HP	1.16	-36.76	0.86	0.14

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. Data range [-366722.3 , 1027438.0]

2) **Detecting**

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

3) **Repairing**

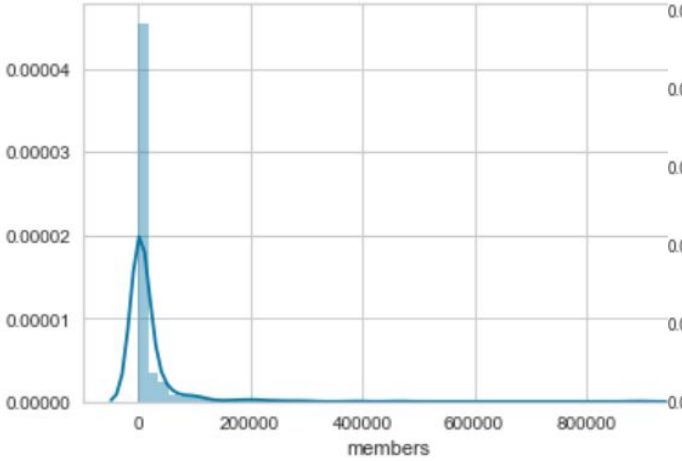
   Using savgol_filter to filter the data from the noise
   Casting
   Drop

GN - Error in 30% - savgol_filter

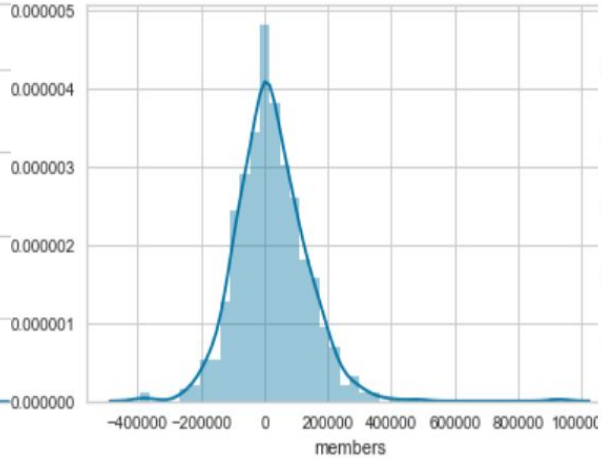
	Before		After	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	0.68	-2.68	0.97	-0.01
Lasso With HP	0.68	-2.68	0.97	0

Members Attribute Before and After Filtering - 30%

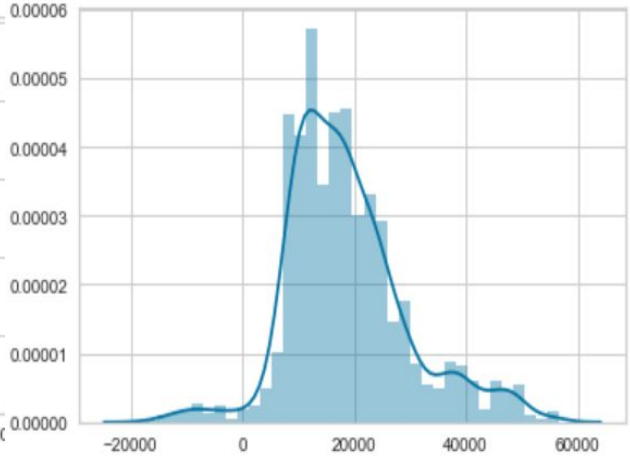
Before Applying GN



After Applying GN



Filter-Gaussian Noise

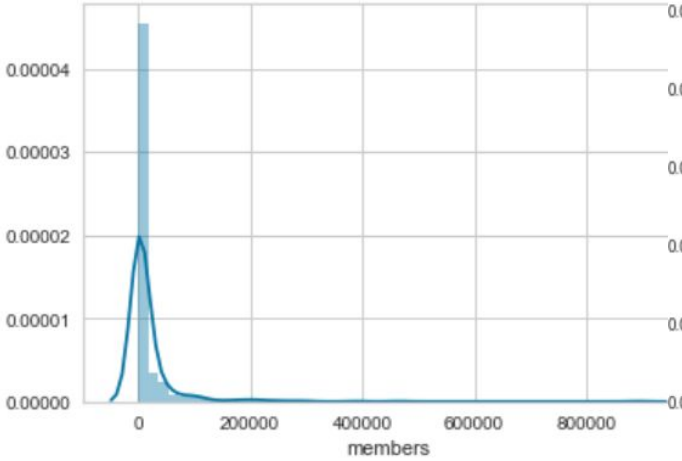


GN - Error on 30% - Casting Data Range

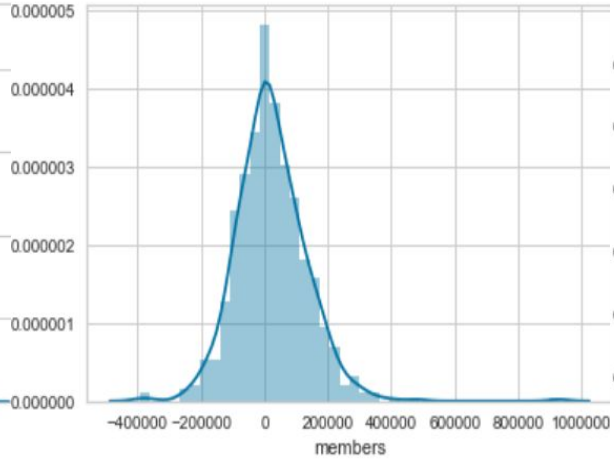
	Before		After	
	MSE on test dataset	Var on test dataset	MSE	Var on test dataset
Lasso Without HP	0.68	-2.68	0.98	0
Lasso With HP	0.68	-2.68	0.97	0.02

Members Attribute Before and After casting the data range- 30%

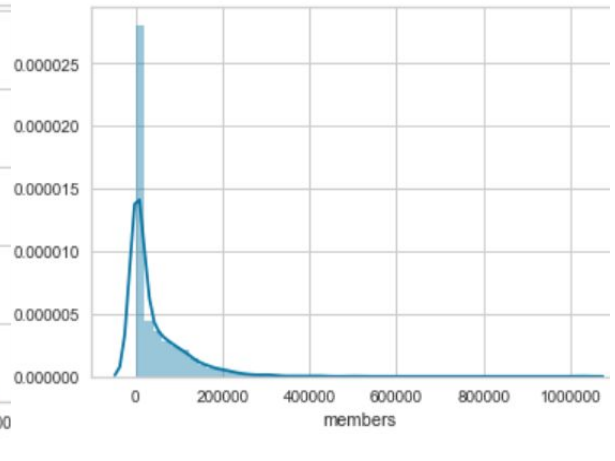
Before Applying GN



After Applying GN



Repairing GN

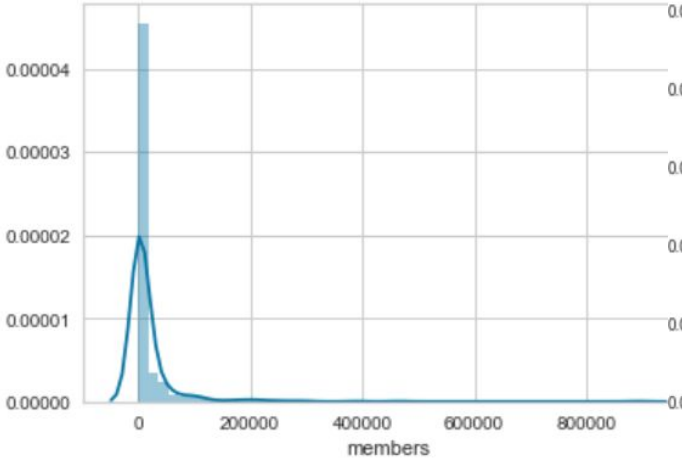


GN - Error in 30% - Replace Negative values By 0

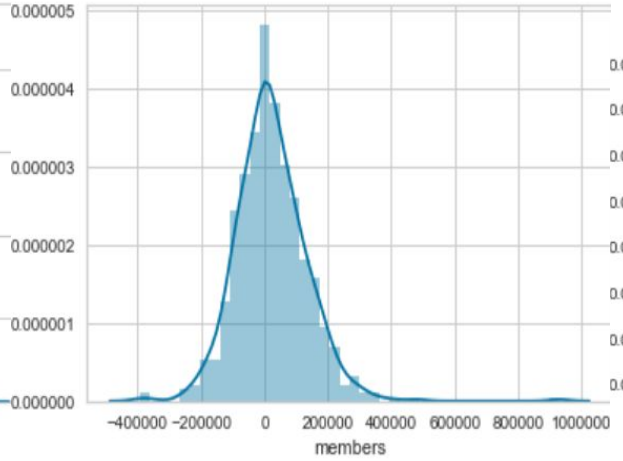
	Before		After	
	MSE on test dataset	**Var on test dataset**	**MSE**	**Var on test dataset**
Lasso Without HP	0.68	-2.68	0.93	0
Lasso With HP	0.68	-2.68	0.93	0

Members Attribute Before and After Replacing by 0 - 30%

Before Applying GN



After Applying GN



Repairing GN