Capstone Project - 2

Sales Prediction : Predicting sales of a major store chain Rossmann

Presented by:
Mrugesh Patel

Let’s sale some drugs!!

1. Problem Statement
2. EDA and feature engineering
3. Feature selection
4. Data preparation
5. Applying different models
6. Model selection

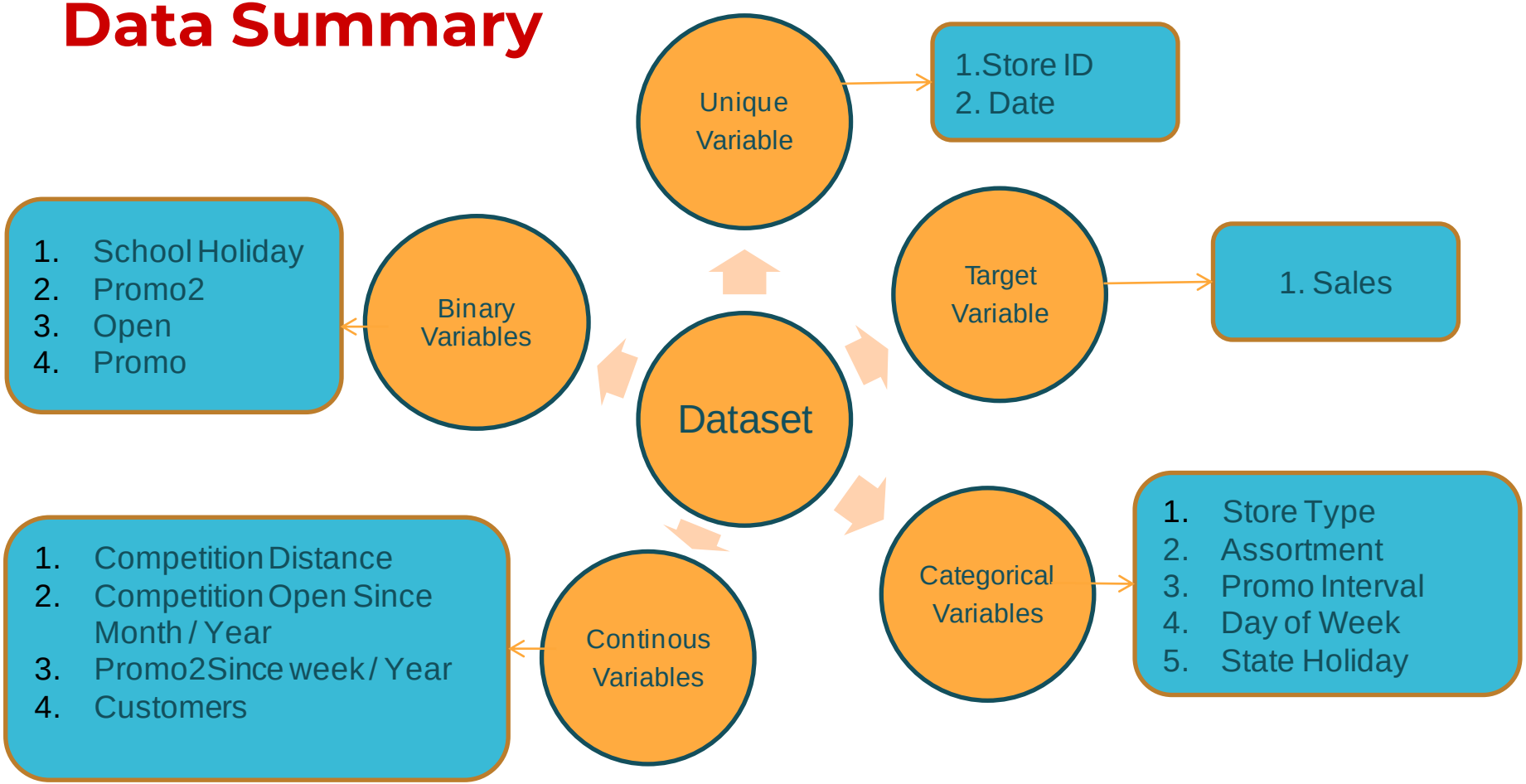
Problem Statement

- Rossmann Store managers are tasked with predicting their daily sales for up to six weeks in advance in order to manage the inventory and drug demand.
- The main objective of the study is to forecast the sales of the test data with at least 95% accuracy of the metric value chosen, which would help the store managers to predict the future sales of the store up to six weeks.

Data Pipeline

- **Data exploration and Reading:** Dataset summary and description, understanding features and target variable
- **Data scrubbing-1:** Removing columns having large amount of Null values
- **Data scrubbing-2:** Encoding of categorical variables and Data imputation of null values, correcting and changing unusable data, feature selection
- **EDA:** Data visualization on features selected from Data scrubbing
- **Data preparation:** Transformation of features and target variable and preparing train and test datasets
- **Creating a Model and Model Selection:** Creating different models with different machine learning algorithms, and selection of best model based on valid metric score using cross validation

Data Summary



Data Reading and Exploration

- Two Datasets – Historical dataset of sales & supplement information
- Common columns Store ID
- More than 1 million historical sales data
- Considerable Null values in CompetitionSinceMonth / Year & Promo2SinceWeek / Year / PromoInterval columns

Data Scrubbing

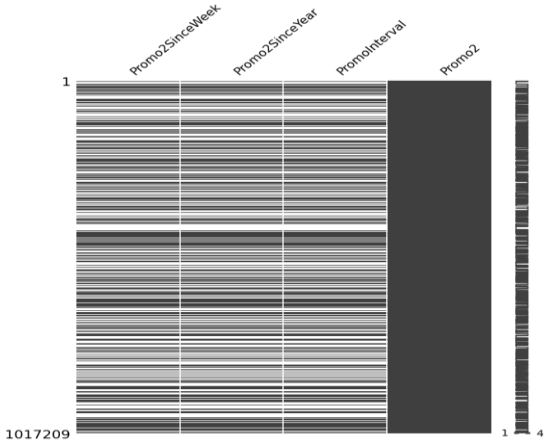
- **Null Value Treatment**

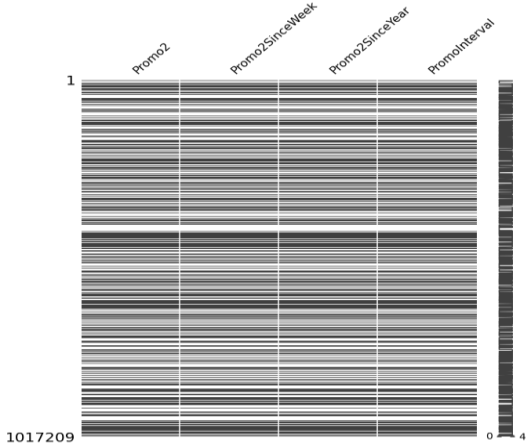
➢Over 50% values are Null in PromoSinceWeek / Year / PromoInterval

➢ Pattern in the Missing values

➢ Assumption : Values are missing where Promo2 is zero, i.e. No Promotion

➢Imputation with 0





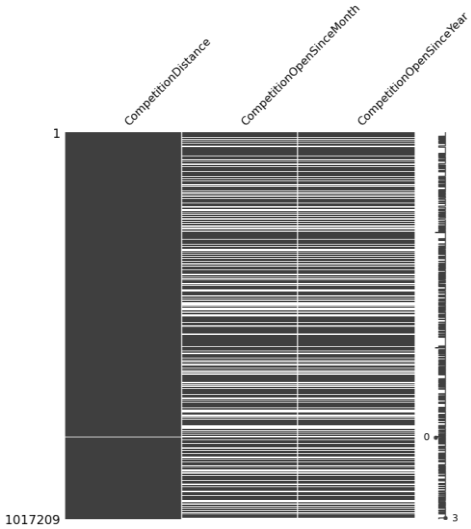
Data Scrubbing

- **Null Value Treatment**

➢ Pattern in the Missing value in CompetitionSinceMonth/ Year

➢ Over 30% values are Null, so columns are dropped

➢ Competition Distance imputed with 0, assuming there is no competition



- **Encoding**

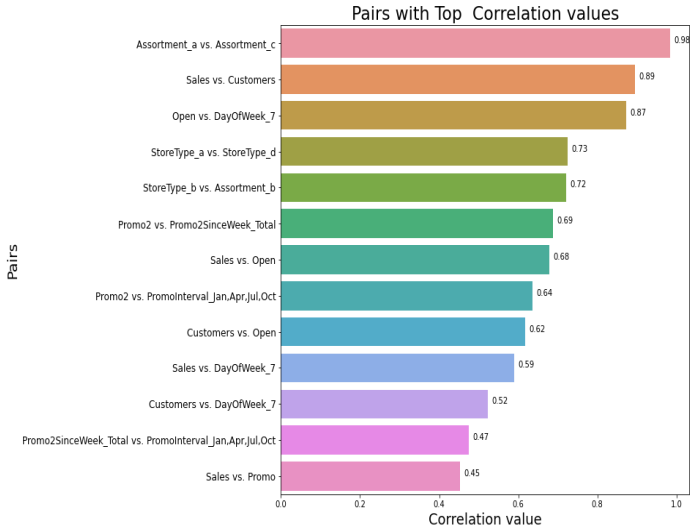
➢ 5 Categorical variables, State Holiday, Assortment, Store Type, Day of Week and Promo Interval

➢ No particular weightage for any categories for each variable

➢One hot encoding

Feature Selection

- Correlation plot for finding highly correlated features
- Avoid multi colinearity
- Top pairs with highest correlations in a bar plot

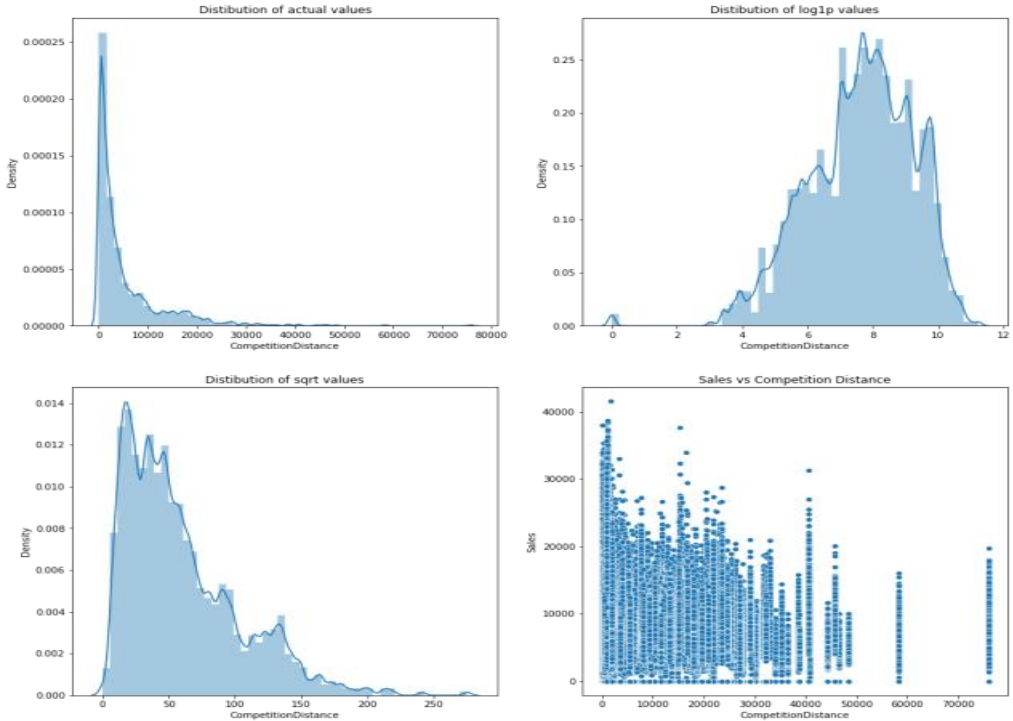


As we can see that The features with high correlation values are,

- Assortment_a
- Assortment_c
- DayOfWeek_7
- Promo2

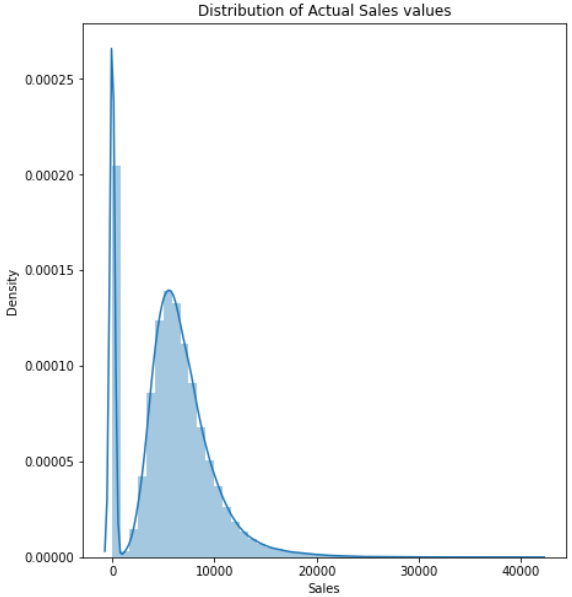
Data Visualization

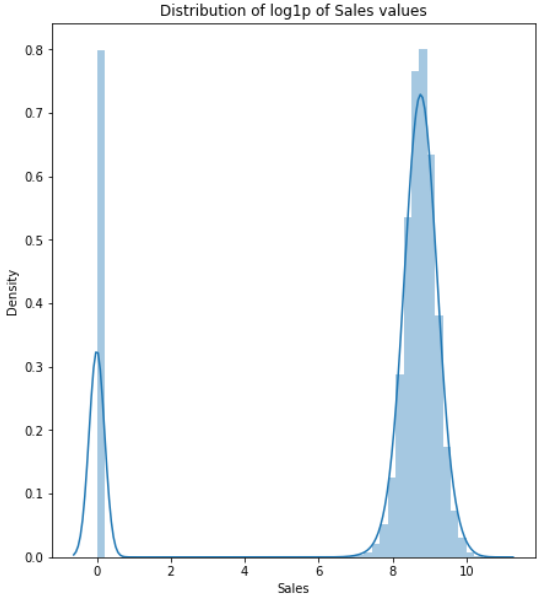
❖ Competition Distance

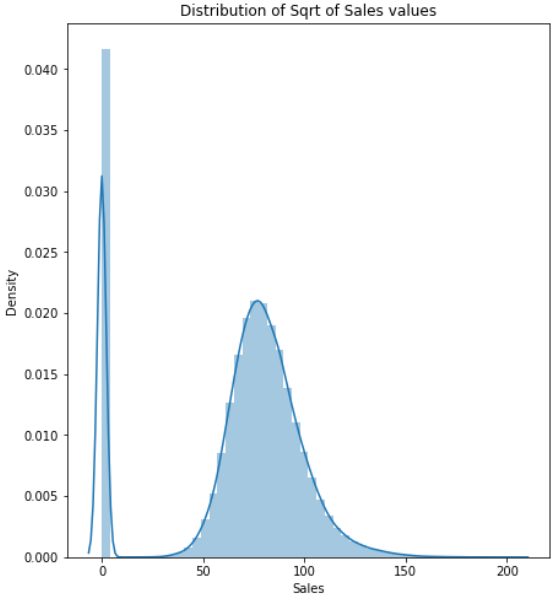


Data Visualization

❖ Sales

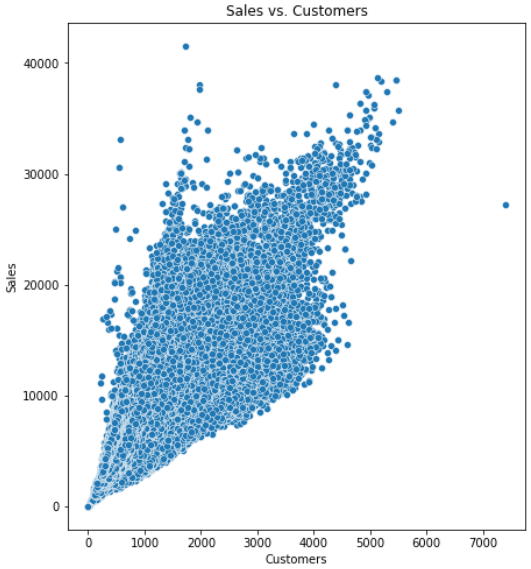


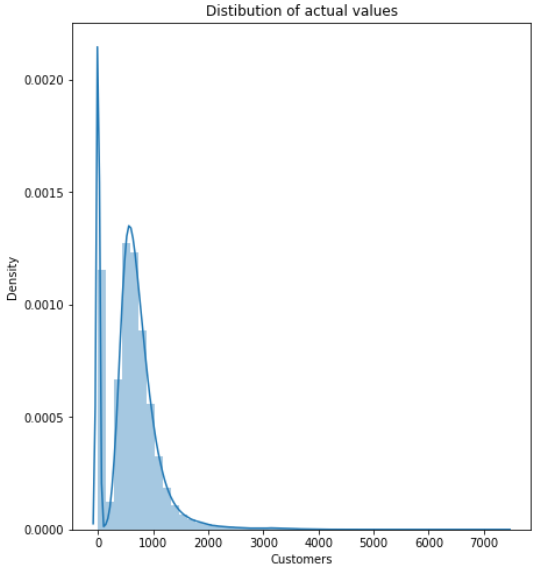


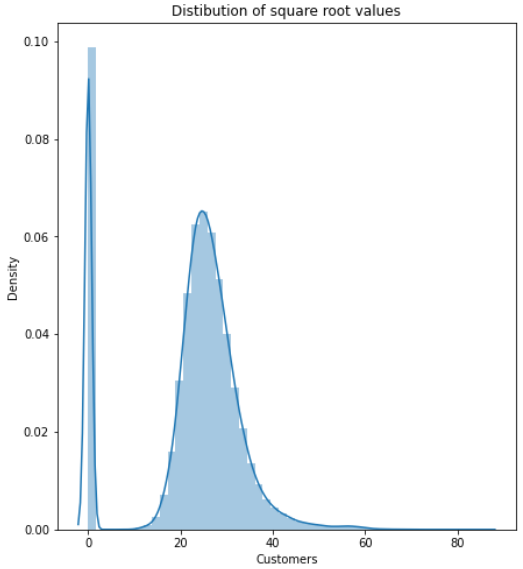


Data Visualization

❖ Customers

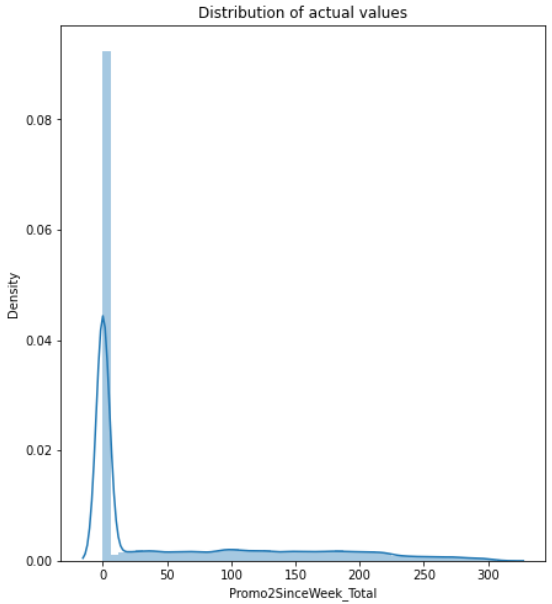


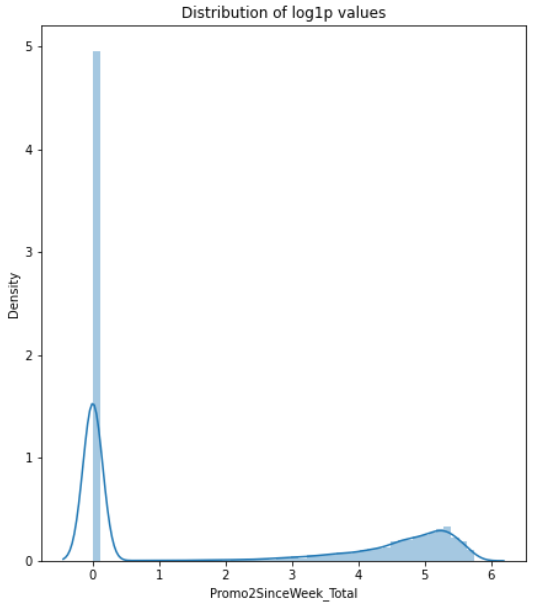


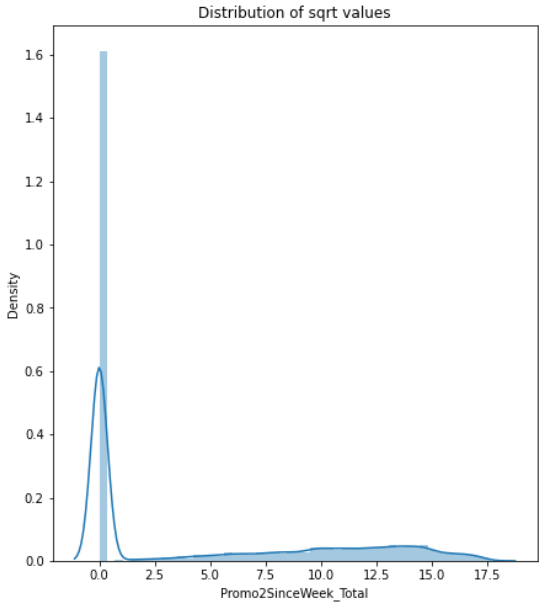


Data Visualization

❖ Promo Since Week (Total)







Data Preparation

❖ Transformation and Standardization

- Some Numerical features as seen before, distribution is not close to Normal distribution
- Transformation of such features required to facilitate the training a better model
- Sales values were transformed to square root values
- Customers and PromoSinceWeekTotal values were transformed with square root values
- CompetitionDistance values were transformed with log1p transformation
- Finally all feature values were standardized before sending the dataset for training, because standardization is very important to achieve uniformity in the dataset in order to have the better model.

Models experimentation

- Base line model : - LinearRegression()
- Regularization using Lasso() and Ridge() with GridSearchCV()
- SGD() and XGBRegressor() with with GridSearchCV()
- KNeighborsRegressor() with GridSearchCV()

❖ **Observations:**

- Baseline Model produced 0.92 r2 score and 1080 RMSE
- Regularization produced similar results and Cross validation for tuning hyper parameters did not improve the model
- Tree based models XGBRegressor performed poorly
- SGD model also produced similar results as Baseline

Results

	MSE	*RMSE*	*R2*	*Adjusted R2*
Linear Regr.	*1.17E+06*	*1081.9*	*0.9205*	*0.9205*
Lasso	*1.16E+06*	*1081.4*	*0.9206*	*0.9206*
Ridge	*1.16E+06*	*1081.4*	*0.9206*	*0.9206*
SGD	*1.17E+06*	*1085.3*	*0.92*	*0.92*
knn	***7.20E+05***	***850.2***	***0.9509***	***0.9509***

❖ Knn model, Hyper Parameters

- 'algorithm': 'auto‘
- 'leaf_size': 30,
- **'metric': 'minkowski',**
- 'metric_params': None
- 'n_jobs': None,
- **'n_neighbors': 5,**
- 'p': 2,
- 'weights': 'uniform'

Future Scope

- Missing values for Competition since Month/Year can be imputed using some machine learning algorithms using other features like knn or svm.
- More models like svm or tree based regressor can be used.

Challenges Faced

- There were 2 datasets with different shape and features
- Promo Since Week / Year not in the usable format
- High Computation time for knn model because of large dataset

Conclusion

- Baseline model Linear Regression produced 0.92 r2 score and 1081.9 root mean squared error. And all three lasso, Ridge and SGD regressor produced similar results as baseline model. Gridsearch CV to tune hyper parameters did not improve the score much. Finally knn model produced 0.93 r2 score initially, and using grid search cross validation to tune hyper parameters, the final best r2 score using knn model was 0.95, and therefore chosen as the final model.
- So after experimenting with multiple algorithms and building multiple model, model with knn-algorithm was picked, as it has the best r2 score among all the models with following metrics:

➢R2 : 0.95096
➢Adjusted R2 : **0.95095**
➢MSE: : 722845.913
➢RMSE: : **850.203454**

Thank You