Project Writeup - Predicting Car Prices

Objectives

Below is the list of questions that we are looking to answer as a final outcome of this Machine Learning project:

- How to develop machine learning model using the K-Nearest Neighbors algorithm.
- How to verify and validate the machine learning model using the K-Fold Cross Validation algorithm.
- How these two separate approaches can be implemented to mitigate the mode fitting and improve model performance?
- What are the effects of underfitting (high bias) and overfitting (high variance) over the model predictability.
- How to apply some of the data cleaning, transforming and feature selecting techniques?
- How feature selection from the training examples impacts the overall predictability of the model?
- How model unfamiliarity with the training set and cross-validation/external data set impacts its predictability performance?

Goal Significance

Why does the list of objectives above matter? What benefit we could derive by understanding the K-Nearest Neighbors and K-Fold Cross Validation algorithm? Below are some the goals that we can enlist. Exercising this technique will help us:

- To determine the impact of model performance using unscrutinized data with ample features Vs. scrutinized data carrying select highly significant features.
- To visualize the effect of underfitting and overfitting with relevance to number of clusters / type of algorithm used / level of data scrutinization etc.
- To understand the effect of number of clusters over the model predictability.
- To identify and remove features that we don't want to use in the model.
- To transform features into the proper format, e.g. scaling of numerical data, fill-in the missing values etc.
- To verify the impact of the above actions by check on the model predictive errors with results comparison.

Data

Data Source

The project uses technical information about various cars from the UCI Machine Learning repository. The repository is available at: https://archive.ics.uci.edu/ml/datasets/automobile

The data used for this project can be downloaded from: https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.data

Data Lists

The data is available in the .csv file format. File name is: `import-85.csv`

Project Writeup - Predicting Car Prices

Data Sampling Methods N/A

Data Extraction Details

The details about the data continuation is available at:
http://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.names

Model

How was the model created?

- Data cleaning, feature scaling, fill out the missing values of useful features with column mean.
- K Nearest Neighbors method is introduced to develop univariate and multivariate models with single and more number of clusters
- Feature segregation to scrutinize the data for more focused analysis
- Kfold Cross Validation algorithm is used to verify and improvise the model performance
- Multiple models and analysis are conducted using the Sklearn packages

Why only this model?

K Nearest Neighbors is very effective and widely used classifier for supervised and unsupervised learning to do predictive analysis as well as data compression and clustering.

K-Fold Cross Validation is very popular algorithm that randomizes the training and cross validation sets within the data set to verify the model predictability over the unknown data. The reliability of the Model gets enhanced when its generalized over the external data can be better checked and confirmed by trying out over the cross-validation set.

Highlights of the code.

Software packages used:

- python
- pandas
- numpy
- math
- Matplotlib.pyplot
- Operator –
  itemgetter
- Sklearn –

  KFold
  mean_absolute_error
  KNeighborsRegressor

Overview:

- Read the data and form the data frame & get familiar with the available data

Project Writeup - Predicting Car Prices

- Select features pertains to the continuous data values based on the info from the source
- Clean the data for special characters
- Convert the data frame data into the same data type for regression
- Clear null values from the training features and target column
- Fill out the missing values of the features with the average values of the column data
- Feature scaling to avoid skew distance matric towards bigger values of the target column
- Develop Univariate and Multivariate models using:
    - K-Nearest Neighbors Regression
- Check and improve model performance by using:
    - KFold Cross Validation
- Display and review Univariate and Multivariate model error matrix - for both approaches
- Visualize the results by plotting the Univariate and Multivariate model errors – for both approaches
- Provide conclusive comments about the results

Model Validation Details.

- Kfold cross validation method used to verify the model predictability over the unknown data.
- Cluster counts and data features were kept uniform to ascertain the proper verification of model performance.
- Model outcomes were closely compared to ensure that the error matrix remains in parity under the both approaches.

Justification for the meaningfulness of the model outcome.

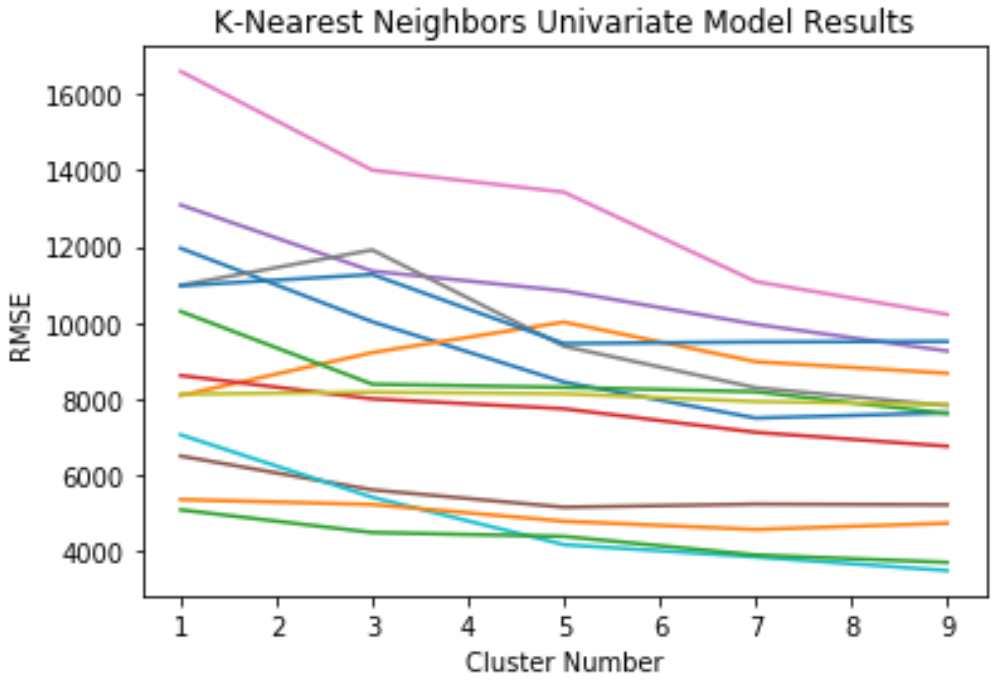
Please refer to the technical explanation of results for the data scientist audience below.

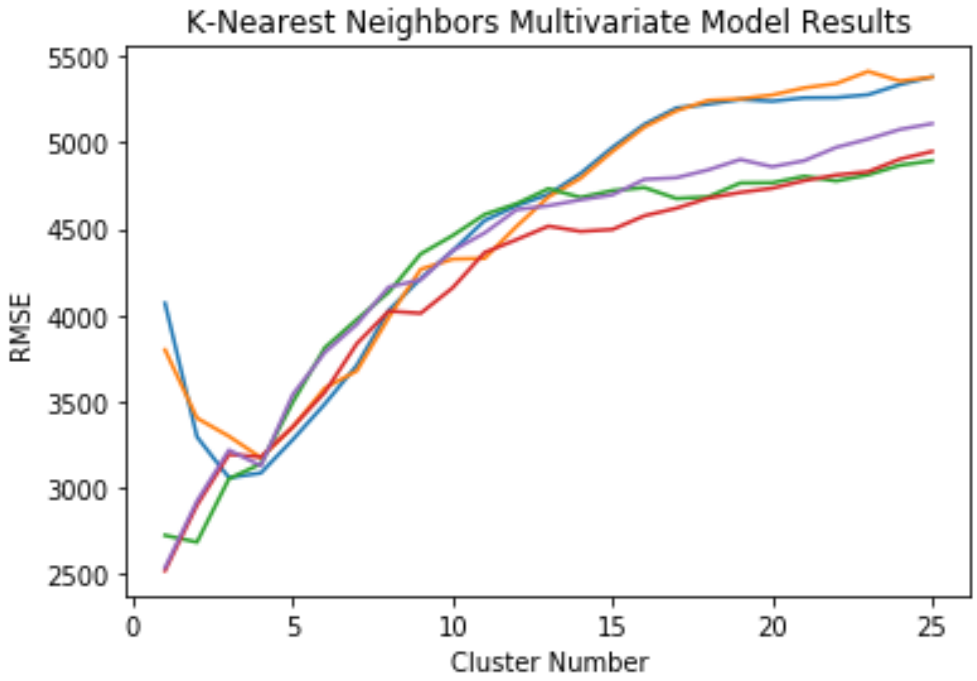
Project Writeup - Predicting Car Prices

Results

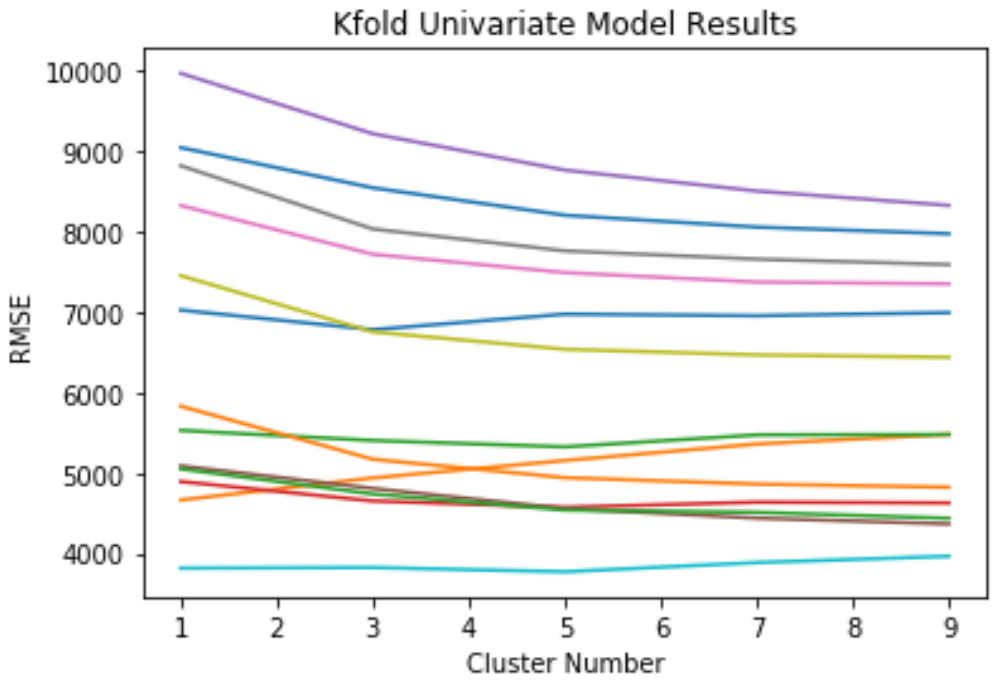
Visualize the results.

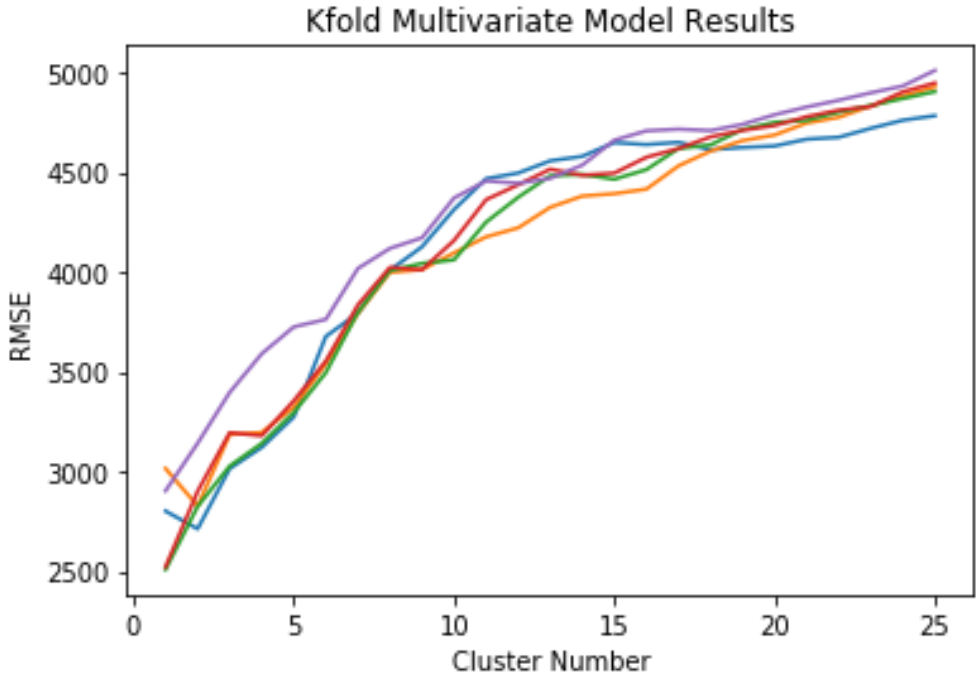
- **Plots for Error Matrices**





Project Writeup - Predicting Car Prices





Project Writeup - Predicting Car Prices

Explain the results in a simple, concise and easy way. (non-technical audience)

- In case of unscrutinized data with lots features, the model performs poorly. However, its predictability improves as we introduce more clusters but the results would show high variability.
- The model performs better if the data is scrutinized by eliminating the unnecessary features and focusing only on the top few highly significant features. The variability of the results also decreases.
- Model performance improves significantly if it gets trained on one set of data and tested on another set of independent/external data.

Explain the results in the most technical way. (technical, data-scientist audience)

- The univariate model of K-Nearest Neighbors shows higher RMSE for low k-values which decreases with the increase in the number of clusters.
    - This is due to the fact that the presence of so many features with low k-value pose a high-bias low-variance scenario that result into an increased RSME.
    - With increase in more clusters force the learning algorithm to better mapping of the nearest neighbors that helps to reduce bias and increase variance. This is reflected in the reduction of prediction errors (RMSE).
- Changes in errors become steady and less intense in K-Fold cross validation to K-Nearest Neighbors. This reflects the effect data due:
    - Shuffle of training and test set in to multiple folds every time, and
    - Nature of the training examples/data points
- In case of multivariate models, top 5 best features are chosen based on the RMSE matrix from the univariate multi-clusters model:
    - The errors are low at the lower number of clusters due to low bias and high variance in the predictions.
    - The errors steadily increase as the number of clusters increases due to increase in bias and decrease in the variance of the results.
- As clearly shown in the graph, the K-Fold cross validation has reduced not only the overall error values but also the variability of the model predictions for the top 5 best features.
- The K-Fold Cross Validation has significantly improved the model performances. This is reflected by steady decrease in the error matrix values as shown below:

Model Type	Change in RMSE by K-Fold Cross Validation
Univariate Model	-6.3%
Multivariate Model with Multi-Clusters	-21.3%
Multivariate Model with top 5 best features	-3.8%

Project Writeup - Predicting Car Prices

Conclusion

What we learn from this outcome. (non-technical audience)

We can summarize our findings from the project outcome as:

1. Reducing the data noise by eliminating the data gaps,
2. effective data cleaning with proper understanding of the data,
3. filling out the missing data properly feature mean values,
4. careful selection of various features of high significance value,
5. reasonably increasing the number data points/clusters, and
6. proper model validation

   can help to improve the predictive model's accuracy for care sale price.

Technical significance of the results. (technical, data-science audience)

- The reduction in data noise helps to outcast the outliers and eliminate resulting errors in the learning algorithm.
- Understand the data, cleaning the data keeping in mind the target outcome, fill out the missing values with appropriate feature value and careful selection of proper features having machine learning significance could considerably improve the model predictability.
- In case of unscrutinized model having almost all of the features under consideration, an increase in number of clusters helps reducing the model error for the given data set. The model predictions for different numbers of cluster also show high variability.
- For scrutinized model with select features under consideration, an increase in number of clusters results in to increase in prediction errors. The model outcome shows relatively low variance for different number of clusters.
- K-Fold Cross Validation helps improving the model performance significantly. This is evident by the collective reduction in error matrix (RMSE) by more than 31% while the model was validated over the CV set.

Suggestion for Further Development

How differently you would have done this project, if get a chance to do it again. Why?

I would have tried following one or more options to explore their possible impact on how the model generalization can be improved:

1. Use the model with 10 clusters while selecting the feature significance for ML
2. Select top 7 or more best features and scrutinize the data for more focused approach
3. Introduce more features by utilizing the existing features to overcome high-bias scenario to begin with.
4. Fill-out the missing values with other than the feature mean.
5. Start project as a fresh with some other supervised learning algorithm such as Linear Regression.
6. Utilize other features with object data type for categorical values.

7. Use of other model validation techniques such as Leave-One-Out Cross Validation, Bootstrap Validation etc.
8. Repeat the points 1 – 4 for their different combinations until the RMSE comes down to a much low value.
9. Repeat the points 5 – 8 if opt to carry out the project with some other algorithm and validation technique until the RMSE comes down within the anticipated range.

Your suggestions for someone if s/he wants to further continue this work.

Someone could pick one of more of my suggested seven points above and continue this journey further to achieve much low RMSE value and further improve the model predictability for the car sale price.