Sales Prediction Data Task

By Nickolas Theodoulou

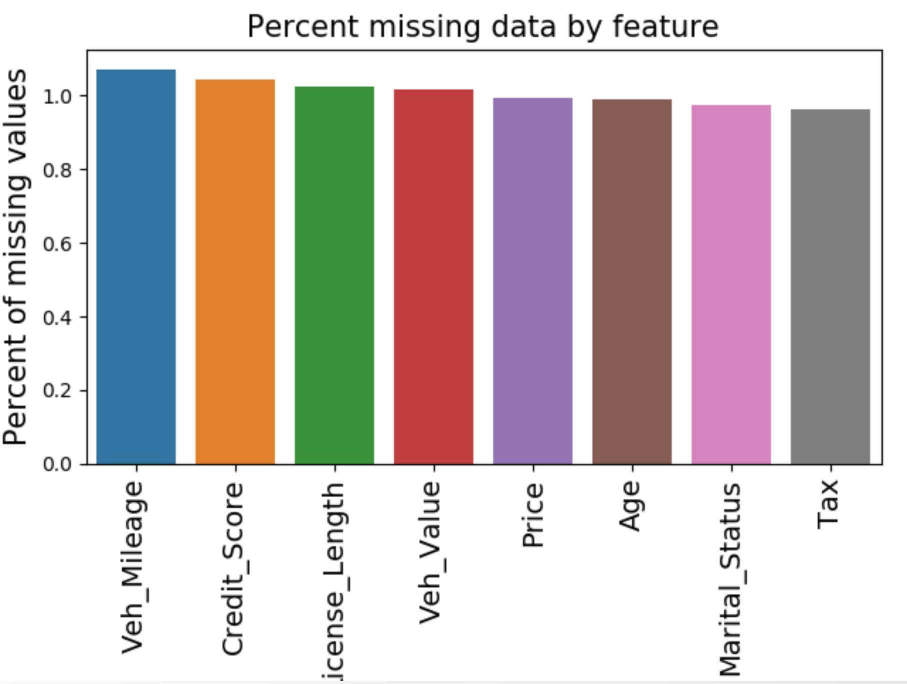
Outline

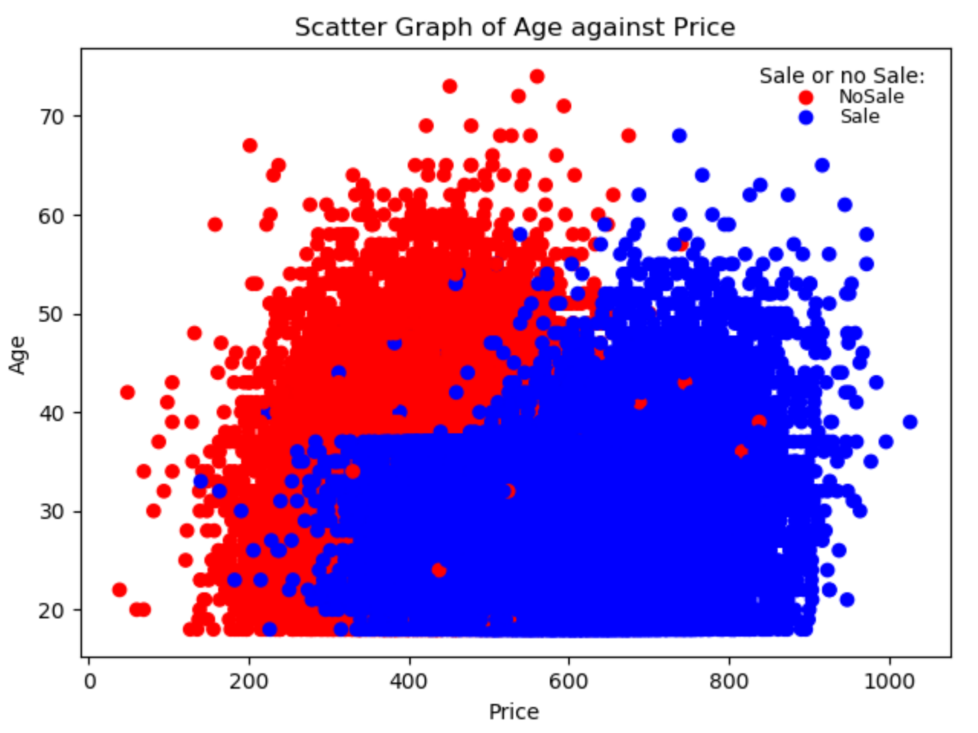
- SVM – Best Model
- Exploration – Interesting Features
- Pre-Processing - Imputation and Transformations
- Modeling – Grid Search and Fitting
- Conclusion – Other Useful Data and Extension

Best Model - SVM

- SVM was found to perform the best using 10-fold cross validation with:
- Mean - 88.3% accuracy, SD - 0.57
- And hyper-parameters: gamma = 1/16, C=10

Data Exploration - Overview

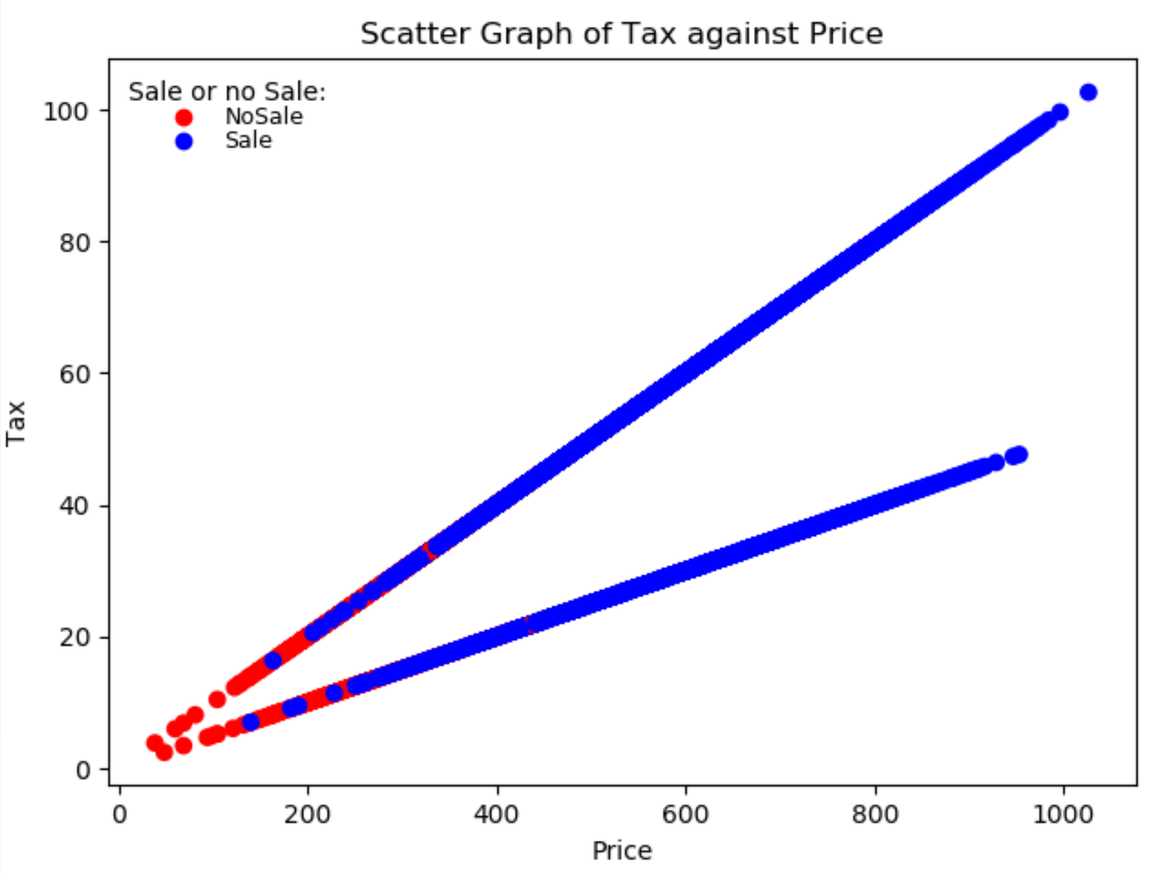




Data Exploration – Correlated Attributes

- Price and Tax were found to be highly correlated from a heat-map

- Linearly dependent following two different equations:
- Tax = Price * 0.1 or Price * 0.05.
- Hence both attributes were used to impute the other using a cutoff value for 0.1 or 0.05.

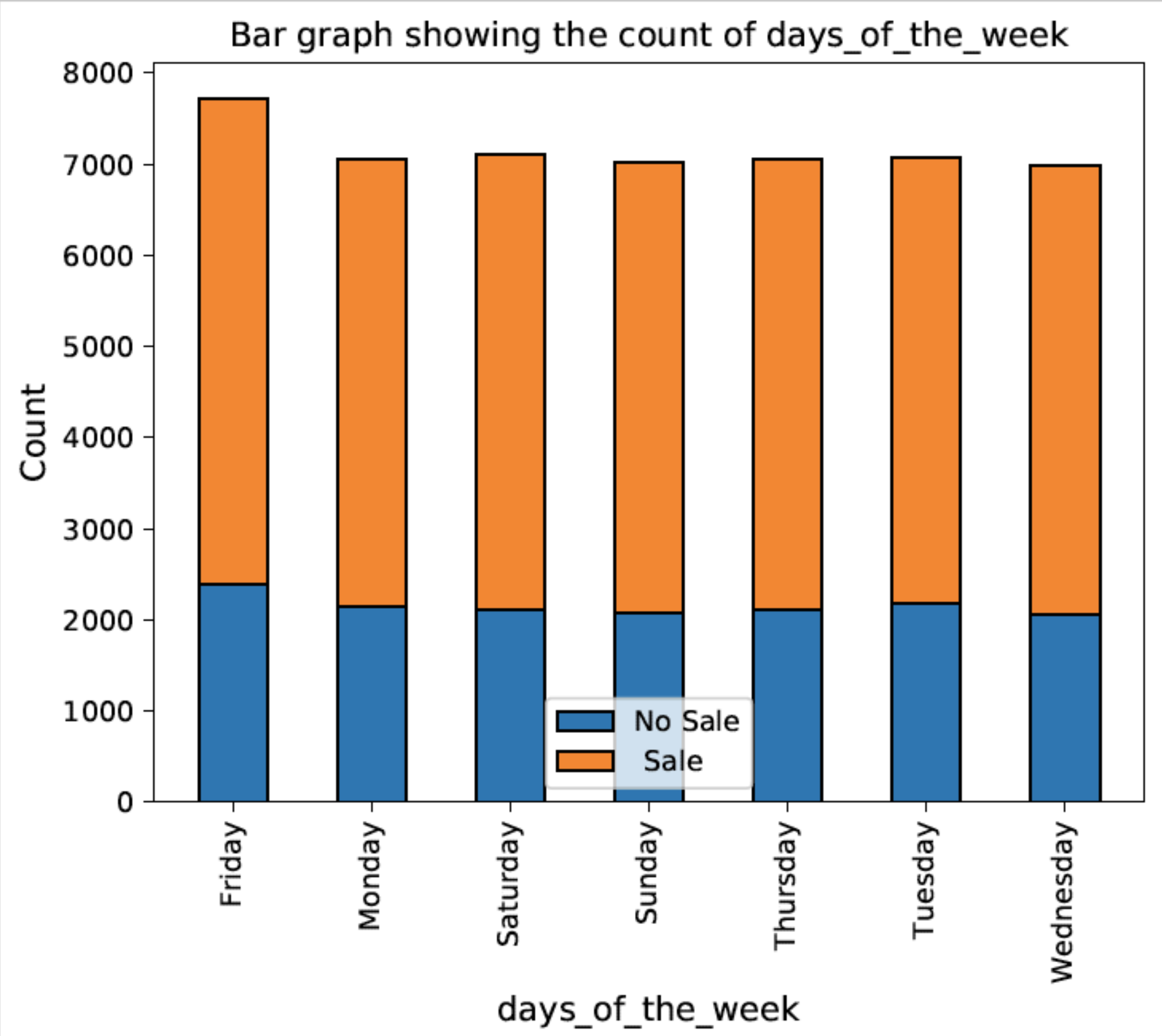


Data Pre-Processing - Imputing

- Attempted to use K-NN to impute attributes – using Package: FancyImpute - large computational cost.
- Hence standard methods such as mean, mode and median were used to impute other attributes.

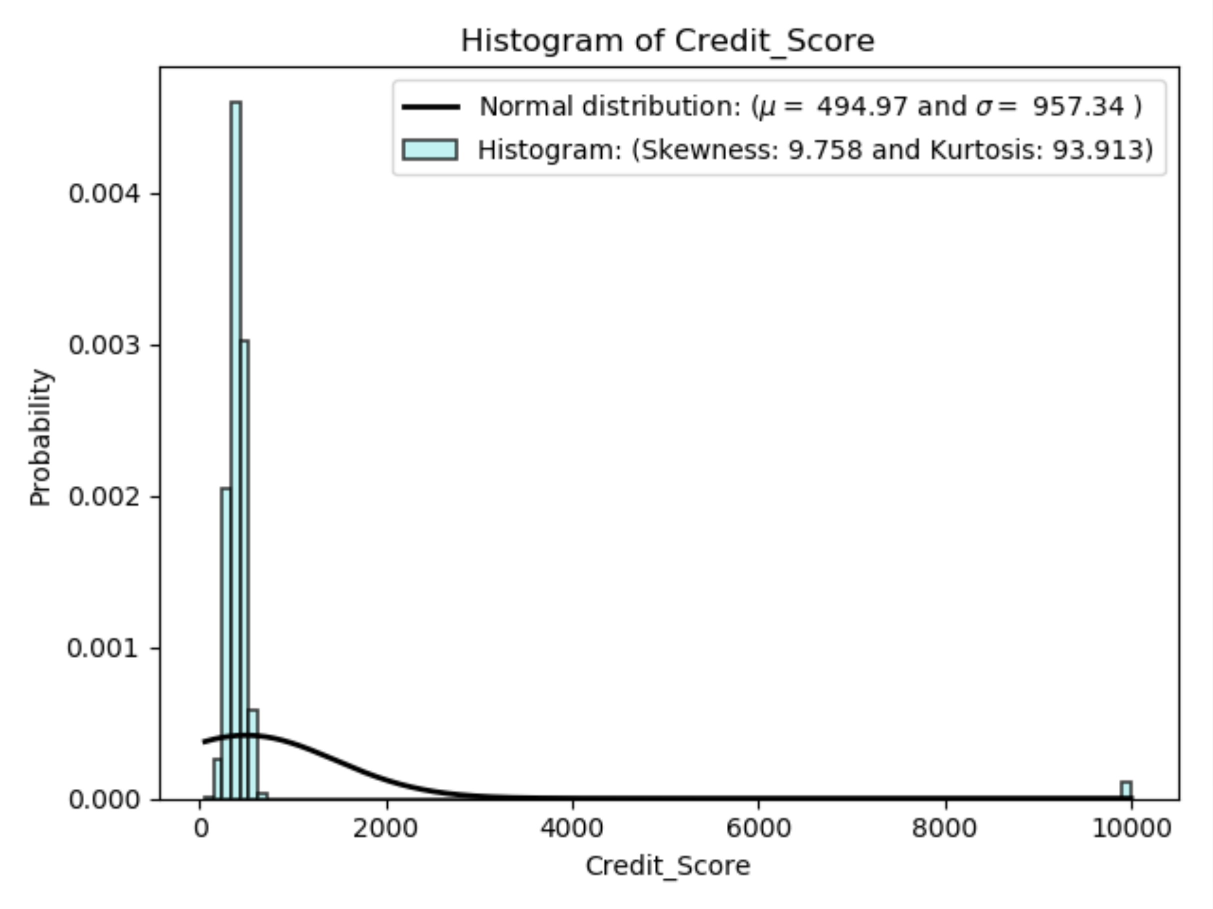
Data Exploration – Feature Engineering

- Days of Week extracted from the attribute Date
- On Friday more sales are made
- Month and Year were extracted - no new information



Data Pre-Processing – One Hot Encoding

- Credit Score had an interesting distribution with customers fitting a bell shape apart from the value 9999.
- Attempted to one-hot encode.



Data Pre-Processing

- Attributes normalised to have a mean of 0 and standard deviation of 1 before fitting.
- No attribute was found to be heavily skewed so log and box-cox transformations were not used.
- One hot encoded Marital Status and Days of the Week.

Data Modeling

- The data was first modeled using K-NN, with k=5, dropping all missing values to see how the data would perform. Accuracy ~ 84%
- After Pre-Processing, grid search and 10-fold cross validation, the optimum value for K was found to be 15 leading to:
- Mean – 87.4% accuracy, S.D- 0.55
- SVM performed better: Mean - 88.3% accuracy, SD- 0.57

Other Useful Data

- Ideal to have more data on customers with no Sale.
- As ‘Date’ was included, Time would have been useful as more sales are likely to occur at certain times of the day.
- Having a customer ID would be useful as the same customer could be inquiring for a quote and this would affect the model.

Conclusion

- Overall the SVM was accurate. Further improvements include:
- Create a method to impute missing values using K-NN
- Fit other ML models after Pre-Processing and compare them
- Stack different models as one could be data points where one model could perform better

Further Questions