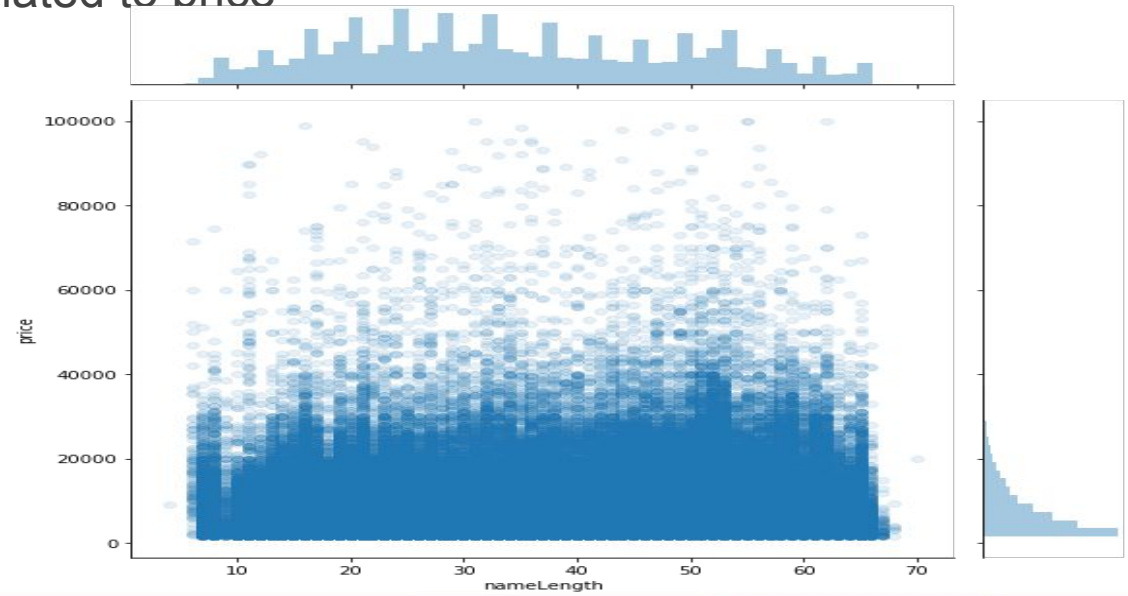
Northwestern

Mid-project Review: Used Car Price Recommender

Sophie Du
May 11, 2019

Highlights

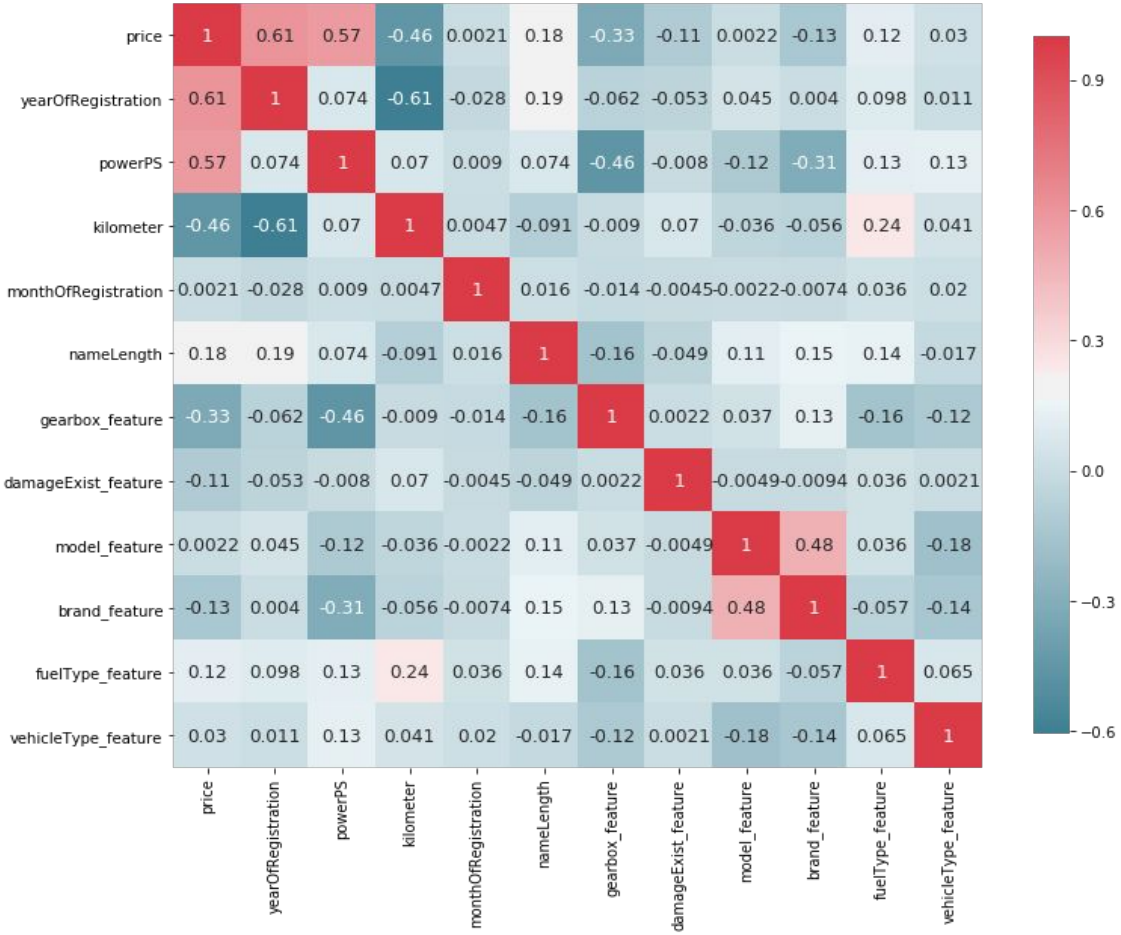
- Completed exploratory data analysis, data cleaning, feature engineering, correlation test and predictive modeling
- Visualization & Interesting findings:
  - Lengths of car descriptions can affect price (15-30 characters gives best result while too long or too short descriptions can negatively influence car prices)
  - Year of registration, horsepower and kilometers are most highly correlated to price



Review progress

- Completed exploratory data analysis by dropping useless or irrelevant columns, checking descriptive statistics of variables and normalizing categorical variables
- Cleaned data (NAs and duplicates removal, outlier detection, fix inconsistent naming, variable selection)
- Performed feature engineering, and checked correlations among all features and how much each feature is correlated to price
- Implemented a predictive model with random forest, used GridSearch to set the optimal parameters for the regressor and trained the final model, and the result shows that 83% variance can be explained by the model
- Checked feature importance in the model

Demo/analysis

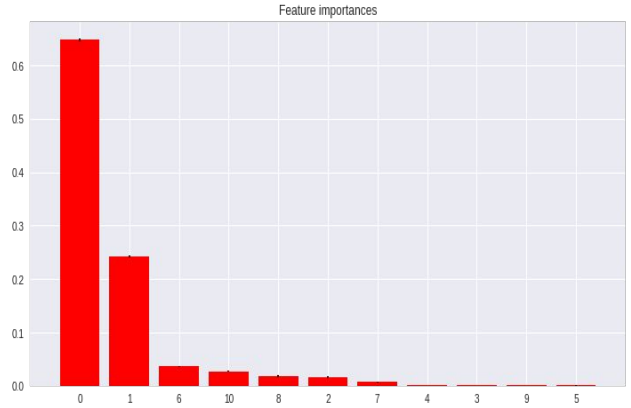


Year of registration, horsepower and kilometers are most highly correlated to price

```
In [21]:
print(gs.best_score_)
print(gs.best_params_)
```

```
0.827770328214
{'criterion': 'mse', 'max_depth': 10, 'min_samples_leaf': 3, 'min_samples_split': 3, 'n_estimators': 500}
```

Sample Prediction Result



Feature importance

Lessons Learned

- Learned how to get a complete plan for a self-managed project
- Tried to manage a large dataset (370k samples) and created new features based on insights for the data and the project goal
- Used scikit-learn preprocessing for categorical variables normalizations and seaborn jointplot for part of the data visualization
- Learned to build part of the pipelines with AWS tools (EC2, S3)
- Handled professional methods of writing docstrings and loggers
- Improved communication skills during peer reviews and discussions

Recommendations

- Finalize model testing and validations and test the model on the full cleaned dataset
- Complete data infrastructure including setting up instances and construct database with code scripts refined in good shape
- Run tests locally and get them pass
- Run tests in a test environment with diverse user inputs (features selected) and improve pipeline performance
- Put code in a scalable production environment
- Write documentation for reference

Thank you for reading!