Capstone Project
Udacity Machine Learning Engineer Nanodegree

Vinicius Viena Santana

March 19, 2021

1 Project Domain Background

In the last few years we have been observing an enormous progress in the performance of computer algorithms for "human-native" tasks: image recognition and classification, voice detection, recognition and translation. The advance of machine learning algorithms are not only automating these tasks but also helping us to better understand how our senses work.

Even tough lot of progress has been made in machine learning for sight and hearing, the same has not been observed for olfaction. Only recently research groups devoted their time to make remarkable progress on it, specially for perfumery, namely: Google Research, IBM Research.

Another remarkable contribution was made by Keller and Vosshall, 2016 5 years ago. They invited 49 subjects to participate in an experiment that involved smelling 476 molecules and registering their perception about its odor characteristics (20 variables). They used this data set to launch the DREAM Olfactory Challenge which aimed to engage machine learning practitioners to develop predictive models to predict subject perceptions of odors given molecular information.

Thus, the aim of this project is developing a machine learning algorithm that can learn to predict odor features of a molecule from its chemical structure using the DREAM Olfactory Challenge data set.

2 Problem Statement

The aim of this project is developing a machine learning algorithm that can learn how to predict odor features of molecules using their chemical structure as input. The data set from of the DREAM Olfaction Prediction Challenge will be used to train the models

3 Data set and Inputs

The data set used for this project is a perceptual data collected from fellows of Rockefeller University Outpatient Clinic. The data set was collected and made publicly available for a data science competition (The DREAM Challenge). A detailed explanation of the data collection procedure can be found at here. I will summarize it here.

They recruited and instructed 49 healthy ethically diverse subjects between 18 and 50 years from the city of New York to participate in the experiment. They collected perceptual ratings of 480 different odors at two different concentrations "high" and "low" for each subject. 20 molecules were tested twice, summing up to 500 stimuli. 338 odors are intended to be used for training (70%), while the data for the 138 odors left is used for validation (69 odors) and the final test set (69 odors).

The training data file has 27 columns. The first 6 columns contain information about the substances (unique identifier and dilution) as well as the subjects (id number). The remaining 21 columns contain subjects perceptual data. Column 7 contains the perceived intensity in a scale between 0-100 where 0 is "extremely weak" and 100 is "extremely strong", column 8 a numerical evaluation of "pleasantness" in a 0-100 scale where 0 is "extremely unpleasant" and 100 is "extremely pleasant". Columns 9-27 contain data in which subjects matched their perception of how the odor smelled to a standard list of 19 perceptual descriptors (Available in the above website)

4 Methodology

4.1 Evaluation Metrics

The metric I will use to evaluate the model is available in the competition website Here I will build the model in the sub challenge 1. It consists in building a model to predict of odor intensity, pleasantness and odor the 19 odor descriptor matrix. The evaluation metric used in the website is the averaged Pearson coefficient for the 49 subjects of all 69 smells. The scores are named r_{int}, r_{ple} and r_{dec}. Int means "intensity", ple means "pleasantness" and dec mean "descriptors". $\bar{r}_{int} = \frac{1}{49}\sum_{n=1}^{49} r_{int_{69odors}}$, $\bar{r}_{ple} = \frac{1}{49}\sum_{n=1}^{49} r_{ple_{69odors}}$, $\bar{r}_{dec} = \frac{1}{49\times 19}\sum_{n=1}^{19}\sum_{n=1}^{49} r_{dec_{69odors}}$. The final score $Score$ is given by $Score = \frac{\bar{r}_{int}+\bar{r}_{ple}+\bar{r}_{dec}}{3}$

4.2 Feature Generation and Engineering

In order to develop the machine learning model, the data is gathered from this repository which corresponds to the data set described above.

In order to develop the machine learning model, the molecules need to be "featurized" – they have to be transformed in useful numeric representation. There are several ways of doing so. Here I will **Dragon** Molecular descriptors (Todeschini and Consonni, 2009) that were released by the challenge organization. It contains 4884 chemical features for each molecule.

4.3 Exploratory Data Analysis and Pre-processing

Before moving to actually building the predictive modeling, is important to "know" the data set well. Depending on the characteristic of the data set, some modeling strategies are more appropriate than others.

First, missing entries will be dropped, than target and features distribution will be assessed to answer questions like: what are the features data types? Does the features have appreciable variability? Are the target or features skewed? Are the features correlated with each other? Which features have the highest correlation with targets?

4.4 Modeling strategy

Since the time is limited, I will train few algorithms to fit the data: ensemble trees (random forest) and a parametric Kernel Ridge Regressor (KRR).

Due to the feature space dimensionality (4886), parametric regression models may work well if the dimensionality is not reduced and features rescaled. For using KRR, the data will need considerable preprocessing – Rescaling and dimensionality reduction. Ensemble trees models, however, do not have any limitation when it comes to feature space dimension or scale. I thiw work I implemented and compared both to check which perform better.

5 Results

5.1 Baseline and Benchmark Models

The challenge launchers released the metrics of a in-house baseline model which achieved **0.273** final Score for sub-challenge 1. They used random forest and tweaked the hyper-parameters to improve performance.

The competition winner reached the outstanding performance of 0.3418 (way higher than the baseline).

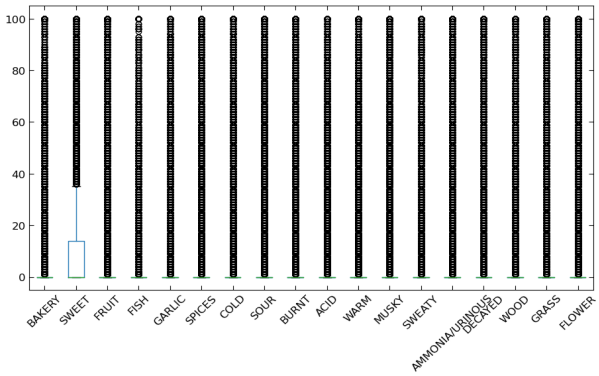
5.2 Exploratory Data Analysis

5.2.1 Target Distribution and Statistics

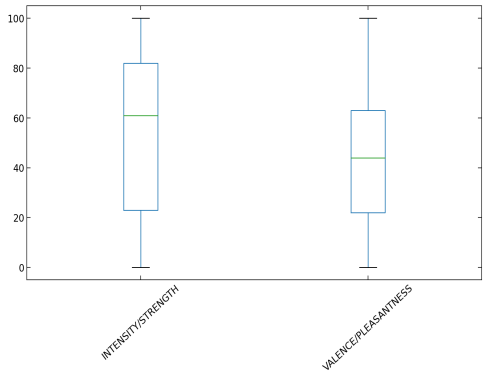
The target data set is very rich and the discussion about it could easily take the whole work. I will provide a small fraction of what it is possible to analyze. The authors of the experiment published a paper discussing several findings and interest patterns in the data (Keller and Vosshall, 2016).

First, a visual representation of data distribution was plotted. The Figure 1 below shows the box plot for 19 of the 21 features. The remaining 2 are plotted separately for better visualization.

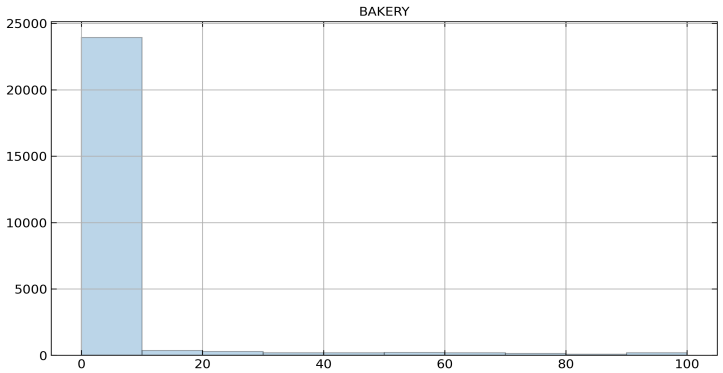
It is possible to notice that, apart from the "sweat" descriptor, most of the subjects choose to use the 0 of the scale. This happened because many molecules had unfamiliar smells according to the study report and of the stimuli that subjects could detect, 70 % were rated as unknown. When confronted with a unfamiliar smell, they could not classify it in any of the 20 possible descriptors. Figure 1c shows how skewed is this descriptor and a similar pattern is found for the others.



(a) Box plot of odor descriptors.



(b) Box plot of intensity and valence.



(c) Histogram of Bakery descriptor

Figure 1: Distribution of target variables

Another interesting aspect of the data set is how each subject report their perception differently for the same odor. The figure below Figure 2 shows the distribution of intensity reported by the 49 subjects for the molecule with identifier 126. It is possible to see that the same molecule can be perceived as "weak" as well as "strong"! The fact that the subjects were "common citizens" that were not trained smell substances may be the cause of such variability. It demonstrates how hard is to describe an olfactory stimuli with numbers consistently.

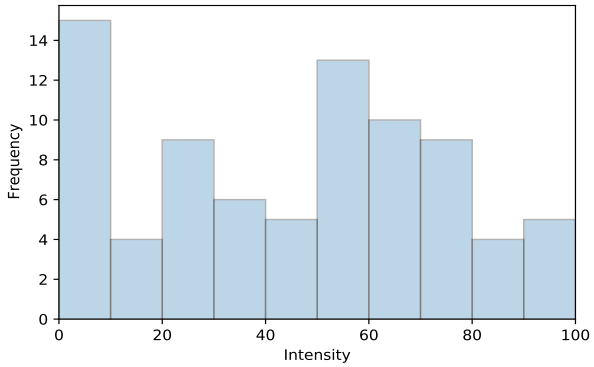


Figure 2: Experienced intensity of component CID 126 for 49 subjects

5.2.2 Features Distribution and Statistics

Before applying regular statistics to the molecular features, it is important to appreciate how rich is of our olfactory system. If we take a look at the molecular complexity distribution of the molecules in the data set, we can see how diverse are the molecules we are sensitive to.

The molecular complexity is a numerical value that reflects the atomic and connectivity diversity of a given molecule – The more distinct atoms it has and more complex are their connections, the higher are their complexity. In the Figure 3 below it is possible to see the distribution of molecular complexity and the structures of few atoms. gray spheres are carbon atoms, red spheres are oxygen atoms, yellow spheres are sulfur atoms and white small spheres are hydrogen atoms. Simple stick is a simple bond and double stick is a double bond. As you we can see, we humans are sensitive to very simple molecules as well as very complex ones. This is fantastic!

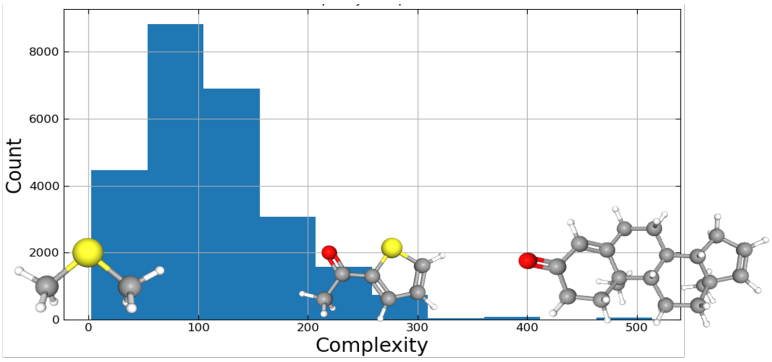


Figure 3: Molecular complexity of molecules in the data set

As there are 4884 features, showing relevant statistics in this document is intractable. The analysis of data types, variability, inter-correlation and symmetry is only relevant for the Kernel Ridge Regressor (KRR) model and not for Random Forest and they will be summarized and discussed below.

In terms of data types, there are 16 features that are binary and the 4853 remaining are continuous. A lot continuous of features have very low variability (variance) and are not useful predictors for the KRR, so they need to be dropped. Similarly, a lot of the binary features are sparse (lot of zeros) and they have to be dropped. Some of the continuous features are highly correlated with each other (up to 95% Pearson coefficient). Thus, they have to be recombined to obtain a new space with a lower degree of correlation for the KRR model – Principal Component Analysis is a suitable method for dimensionality reduction for instance.

5.3 Training and evaluation

5.3.1 Random Forest model

For the random forest model, all "Dragon" features available were used without any preprocessing. One of the advantages of random forest is its robustness concerning features' dimensionality and scales. The sub-challenge 1 (SB1) were divided into 2 tasks. The first task (T1) is to predict odor intensity and the second (T2) is to predict the remaining

targets (valence and odor descriptors). The model was implemented in a ScikitLearn custom estimator in SageMaker.

For **SB1-T1**, a hyperparameter tunning job were run in SageMaker for maximizing $\bar{r}_{int}$ in the leaderboard. The hyperparameters and their search space ranges were: {"min samples leaf":[2, 8], "min samples split": [4,16], "estimators": [30, 70]}. The maximum $\bar{r}_{int}$ was 0.28 for "min samples leaf" = 6, "min samples split" = 16 and "estimators" = 61. The table below shows the performance of the top 3 models.

Estimators	min samples leaf	min samples split	r_{int}
61	6	16	0.280725
63	6	15	0.278514
62	4	15	0.275987

Table 1: Best 3 Random Forests for Intensity. **(SB1-T1)**

After that, the best model was evaluated on the test set using the metric described in subsection 4.1 resulting in $\bar{r}_{int} = 0.3466$ for this model. You can see that the model achieved a much higher performance in the test set. It can be a sign that the data distribution of the leaderboard set is quite different that the test set – It was "easier" for the random forest to predict in the test set.

For the second task **(SB1-T2)** only the random forest with a multitask learner will be implemented due to time constraints. In future works, other algorithms can be tested for (SB1-T2).

For **SB1-T2**, a hyperparameter tunning job were run in SageMaker for maximizing $\bar{r}_{ple} + \bar{r}_{dec}$ in the leaderboard. The hyperparameters and their search space ranges were: {"min samples leaf":[2, 15], "min samples split": [4, 50], "estimators": [20, 70]}. The maximum $\bar{r}_{ple} + \bar{r}_{dec}$ was 0.462 for "min samples leaf" = 12, "min samples split" = 17 and "estimators" = 68. The table below shows the performance of the top 3 models.

estimators	min samples leaf	min samples split	FinalObjectiveValue
68.0	12.0	17.0	0.461973
68.0	3.0	4.0	0.459336
64.0	10.0	41.0	0.456904

Table 2: Best 3 Random Forests for valence and odor descriptors summed. **(SB1-T2)**

After that, the best model was evaluated on the test set using the metric described in subsection 4.1 resulting in $\bar{r}_{val} + \bar{r}_{dec} = 0.527$ for this model. You can see that the model achieved a much higher performance in the test set. It can be a sign that the data distribution of the leaderboard set is quite different that the test set – It was "easier" for the random forest to predict in the test set.

5.3.2 Kernel Ridge Regressor (KRR)

As mentioned in subsubsection 5.2.2, a lot of preprocessing is necessary for using a model like KRR, namely: drop features with low variance and reduce auto-correlation via feature linear combination. For the latter, PCA was used to obtain a combination of the original feature space that could retain 98% of the variability. After PCA, the transformed space (Z) were rescaled with mean and standard deviation (Standard Scaling). After that, 66

features "survived" which is much less than the original space and suitable for a model like KRR.

The figure below shows a diagram of the features preprocessing steps

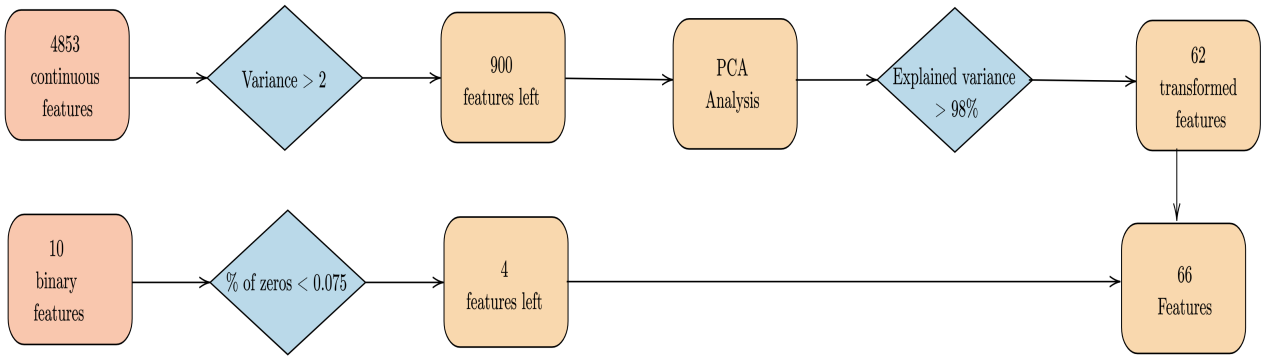


Figure 4: Preprocessing steps

For **SB1-T1**, a KRR with radial basis function were used. a hyperparameter tunning job were run in SageMaker for maximizing $\bar{r}_{int}$ in the leaderboard with 15 jobs. The hyperparameters and their search space ranges were: {"alpha":[0.01, 0.1], "gamma": [0.1, 1]}. These parameters corresponds to the L2 regularization parameter and inverse of standard deviation of radial basis function kernel. The maximum $\bar{r}_{int}$ was 0.226011 for "alpha" = 0.032118, "gamma" = 0.1 . The table below shows the performance of the top 3 models.

alpha	gamma	FinalObjectiveValue
0.032118	0.1	0.226011
0.030319	0.1	0.226011
0.079896	0.1	0.226005

Table 3: Best 3 Kernel Ridge for Intensity. **(SB1-T1)**

After that, the best model was evaluated on the test set using the metric described in subsection 4.1 resulting in $\bar{r}_{int} = 0.3775$ for this model. You can see that the model achieved a much higher performance in the test set. It can be a sign that the data distribution of the leaderboard set is quite different that the test set – It was "easier" for the KRR to predict in the test set. Also, this model performed better than the Random Forest. It seems that the leaderboard set is not a good proxy to assess the performance of the model on the test set. The KRR with a lower performance on the leaderboard test actually performed better on the test set than Random Forest.

5.3.3 Final result

To get the final result, $\frac{\bar{r}_{ple}+\bar{r}_{dec}+\bar{r}_{dec}}{3}$ have to be calculated. Here I will calculate the metrics for two combinations of predictors in each of the tasks.The following list show the results.

1. Random Forest (SB1-T1) and Random Forest (SB1-T2). Score = 0.2912

2. Kernel Ridge Regressor (SB1-T2) and Random Forest(SB1-T2). Score = 0.3015

Note that for both options, it was possible to overcome the baseline model performance developed by the launchers of the competition. However, I could not beat the champion. Still, I would be inside the top 10 competitors.

6 Conclusions

This work addresses and old problem in science which is predictive modeling for olfaction. Here I used a famous competition called the Dream Olfactory Challenge. The data set consists in the registered experience of 49 ethically diverse people for 500 olfactory stimuli (480 molecules).

The main objective of this challenge was stimulating machine learning practitioners to develop models that can predict 21 features of the molecules. The challenge was divided into 2 subchallenges. Due to time constraints, this work only address the first challenge.

The base line model developed by the competition launchers reached a performance of **0.273** while the winner reached **0.3418**. In this work, I developed random forest and kernel ridge regressor models that reached a performance of **0.3015** which would place me in the top 10 competitors (7th place).

References

[1] Andreas Keller and Leslie B Vosshall. "Olfactory perception of chemically diverse molecules". In: *BMC Neuroscience* (2016), pp. 1–17. DOI: 10.1186/s12868-016-0287-2.

[2] Roberto Todeschini and Viviana Consonni. *Molecular Descriptors for Chemoinformatics*. Vol. 41. Methods and Principles in Medicinal Chemistry. Wiley, July 2009, pp. 1–252. ISBN: 9783527318520. DOI: 10.1002/9783527628766. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9783527628766.