Methodology for Recipe Sales Forecasting

Recipe Sales Forecasting

Objective:

Forecast recipe sales for a particular recipe id and year_week

Exploratory Data Analysis:

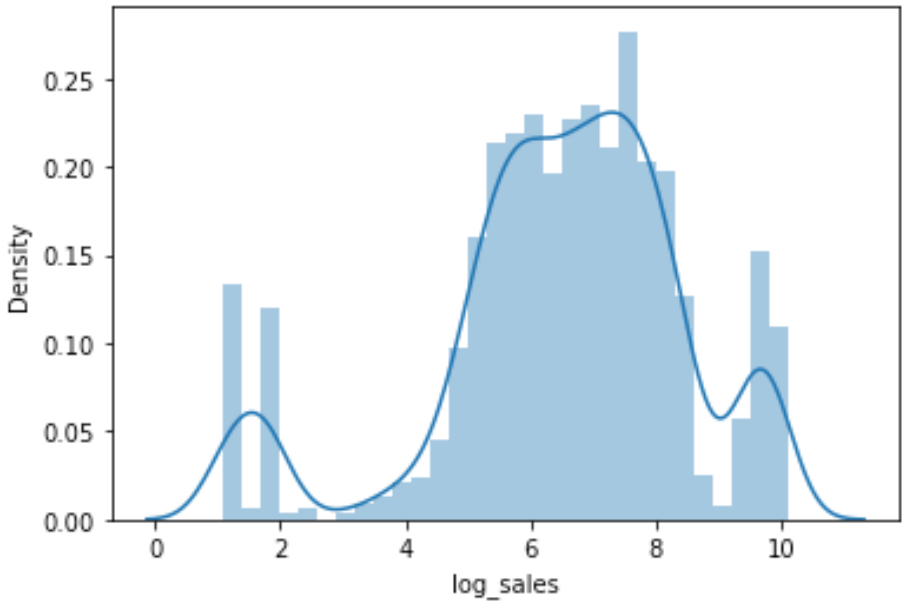
The dataset contains a large number of categorical variables. Variables like recipe_name, description contains text descriptions of the recipes. This information on the recipe can be used for the purposes of recipe sales prediction. This is done using NLP techniques by computing embeddings on each of the text variables that enables them to be grouped together.

Number of categories for Text variables

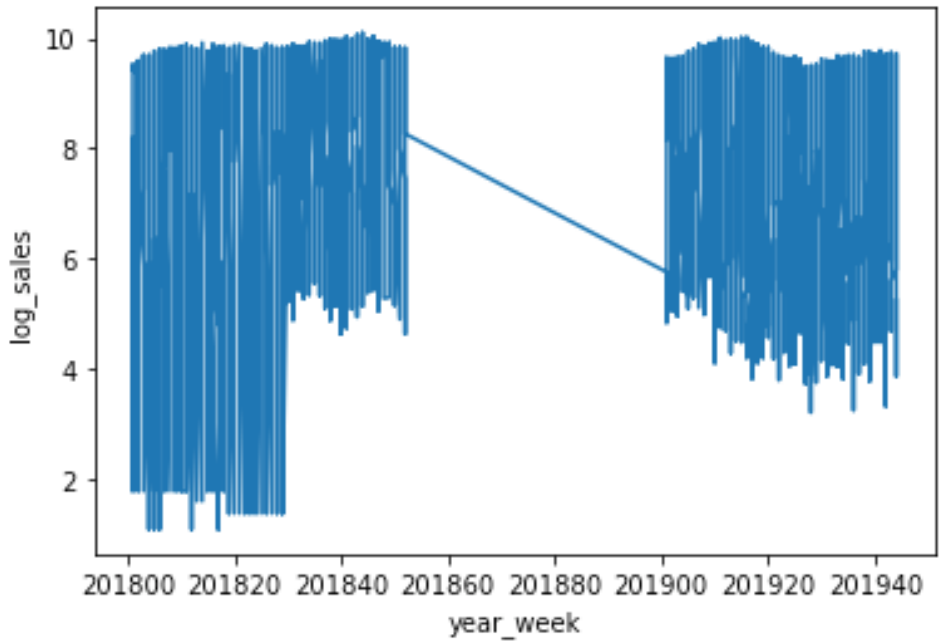
```
recipe_name 931
product_type 4
cuisine 28
description 981
difficulty 2
dish_type 5
is_classic 1
preferences 45
carbs_content 44
dish_types 109
seasons 11
protein_types 35
course_type 2
meta_tags 15
protein_cuts 17
```

Log-sales density plot

Log of sales density appears to be normaly distributed



Log-sales over year_week



Correlation of Continuous variables with Sales

Given below is a correlation heatmap of log_sales with all the continuous variables in the dataset. Year is considered a categorical variable here since there are only two years. Quite

interestingly, Carbs, fats and proteins are highly correlated with log_sales indicating peoples interest in an unhealthy diet.

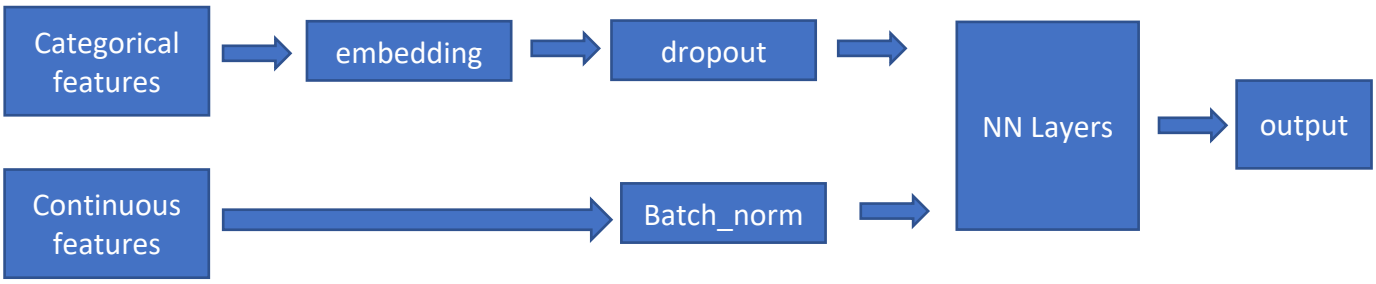
Feature Engineering on categorical variables

Cooking_time: converted to an ordered encoding of from [3,4,5,6,7]
Carbs_content: The '_' in the carbs_content text is removed to create a brief description of carbs_content. For example 'pasta_incl_gnocchi_spatzle' becomes 'pasta incl gnocchi spatzle'.
Protein_cuts: same transformation performed as carbs_content
heat_content: an ordered encoding is performed as follows {nan: 0, 'no_heat': 0, 'optional_heat': 1, 'high':3, 'non_spicy':2}
sales: converted to log_sales inside the fastai library
year_week : transformed into week number and year.

Methodology

The fastai.tabular library is used to train a model and predict sales. The fastai library trains a neural network that takes categorical and numeric variables as inputs and processes them through a neural network. Dropout is implemented on the categorical features and batch normalization on the continuous features

Neural Network schematic Architecture



The fastai library also trains a neural network to learn embeddings on the categorical variables (texts). The dimension of the embedding vector is set to minimum (50, cardinality of the categorical variable). Cardinality of the categorical variable is equal to the number of categories in the categorical variable + 1.

Shape of the embedding vector

	cat_column	embed_shape
0	recipe_name	50
1	product_type	2
2	cuisine	13
3	description	50
4	difficulty	0
5	dish_type	2
6	heat_level	2
7	is_classic	0
8	preferences	22
9	carbs_content	21
10	dish_types	50
11	seasons	5
12	protein_types	17
13	course_type	0
14	meta_tags	7
15	protein_cuts	8
16	year	1

Results

A K – Fold cross validation is performed on the model outputs. We get the following results. The RMSE and R Squared metric below is computed on log sales.

Fold	RMSE	R Squared
1	0.6718	0.783
2	0.5681	0.828
3	0.5523	0.8863
4	0.6122	0.846
5	0.6173	0.8094