Materials Informatics Web Application

Exequiel Punzalan

Deliverables

1. compressed file
    - Source code
        - *webapp.py*: Python code for streamlit app
        - *test.ipynb*: Python code for featurizing data, connecting to MySQL, and training the ML model
        - *app_utils.py*: utility functions
        - *requirements.txt*: package to be installed that were used in the Python environment
    - Instructions on how to setup and run locally
    - Summary of work

Instructions

1. Setting up the Python environment with Bash:

The environment can be installed manually using the following:

```
> conda create -n matminer-env python=3.10
> pip install -r requirements.txt
```

OR install my specific package versions using pip with Python 3.10

```
> pip install -r requirements_specific.txt
```

2. Running the web application locally with Bash:

```
> conda activate matminer-env
> streamlit run webapp.py
```

3. The file test.ipynb contains the code for features, SQL, and the ML model. To examine and/or rerun the code,

- Open test.ipynb
- Change username and password in MySQL section to connect to MySQL locally
- Select matminer-env as the kernel/environment
- Click “Run All” in the ipynb or “Run” select cells

Summary

A web application made in Python was made that obtains materials data from Matbench and uses a machine learning (ML) model trained on the data to predict a material property from composition. Specifically, the Matbench experimental band gap composition dataset was used to predict band gap. The interactive application built with Streamlit allows users to input material composition and display predictions for the band gap.

Dataset

The Matbench experimental band gap dataset was obtained using Python via the matminer package. The dataset contained the composition and experimental band gap as columns.

Technologies Used

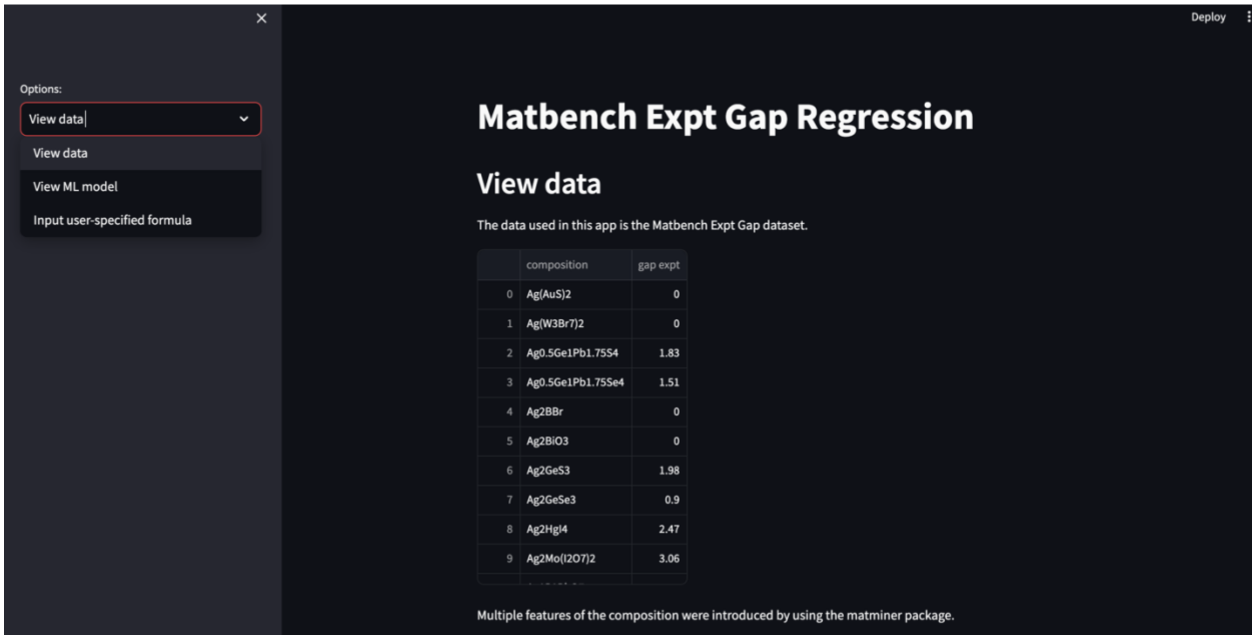
Python was mainly used in this project. The matminer package was used to featurize the data, which were used as inputs for the ML model. The packages sqlalchemy and mysql-connector-python were used to connect and communicate with MySQL. The scikit-learn package was used for training and testing the ML model. The model was saved as a file using joblib. The streamlit package was used for the web application.

Design Decisions

Main design decisions involved featurizing the data, handling data structures, connecting Python with MySQL, selecting the ML model, and organizing the streamlit application.

The matminer package was used to expand the dataset because it included systematic ways to featurize material composition data. The data were mostly handled as pandas DataFrames in Python and converted to tables in MySQL. These two data structures can be inter-converted which allows the ML model to also be trained on SQL tables. The random forest regressor was used as the ML model due to its flexibility and ease-of-use, while not requiring substantial hyperparameter tuning.

The design of the streamlit app mainly involves a sidebar option to view different results (**Figure 1**). The sidebar allows for three different options: viewing the data, viewing the ML model results, and providing user-specified input. The 'View data' option allows the user to look at the original and featurized datasets. The 'View ML model' option allows the user to examine the prediction accuracy of the ML model via regression plots. The 'Input user-specified formula' option (**Figure 2**) allows the user to input composition and obtain a prediction of band gap based on the trained ML model.



Matbench Expt Gap Regression

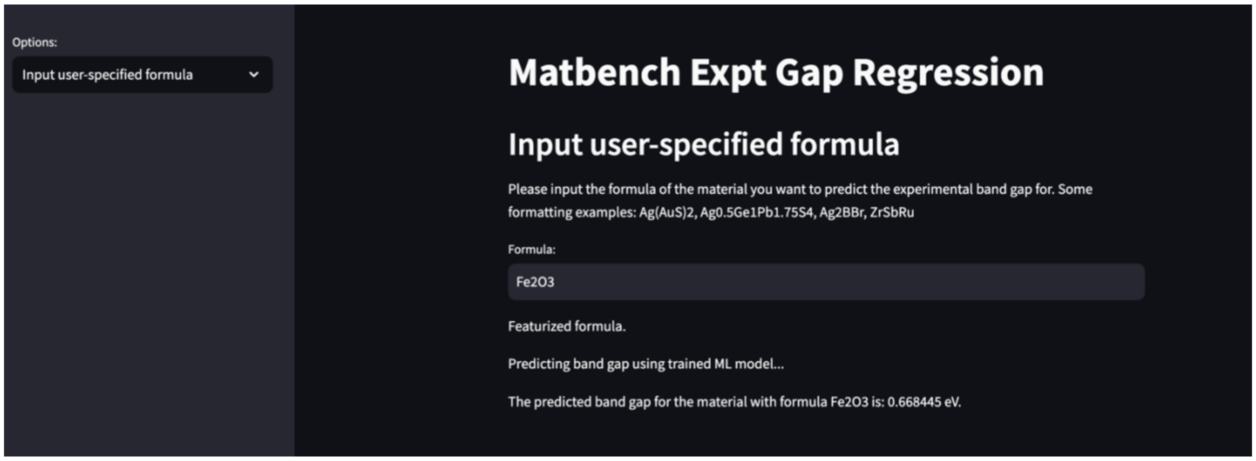
View data

The data used in this app is the Matbench Expt Gap dataset.

	composition	gap expt
0	Ag(AuS)2	0
1	Ag(W3Br7)2	0
2	Ag0.5Ge1Pb1.75S4	1.83
3	Ag0.5Ge1Pb1.75Se4	1.51
4	Ag2BBr	0
5	Ag2BiO3	0
6	Ag2GeS3	1.98
7	Ag2GeSe3	0.9
8	Ag2HgI4	2.47
9	Ag2Mo(I2O7)2	3.06

Multiple features of the composition were introduced by using the matminer package.

Figure 1. Design of the streamlit web application.



Matbench Expt Gap Regression

Input user-specified formula

Please input the formula of the material you want to predict the experimental band gap for. Some formatting examples: Ag(AuS)2, Ag0.5Ge1Pb1.75S4, Ag2BBr, ZrSbRu

Formula: Fe2O3

Featurized formula.

Predicting band gap using trained ML model...

The predicted band gap for the material with formula Fe2O3 is: 0.668445 eV.

Figure 2. Display of prediction for user-specified composition.

<u>Model Performance</u>

Appropriate plots and metrics showing the prediction accuracy of the model are shown in the streamlit application and attached as files in the plots directory. The random forest model performs well on test data with $R^2 = 0.78$ (**Figure 3**).

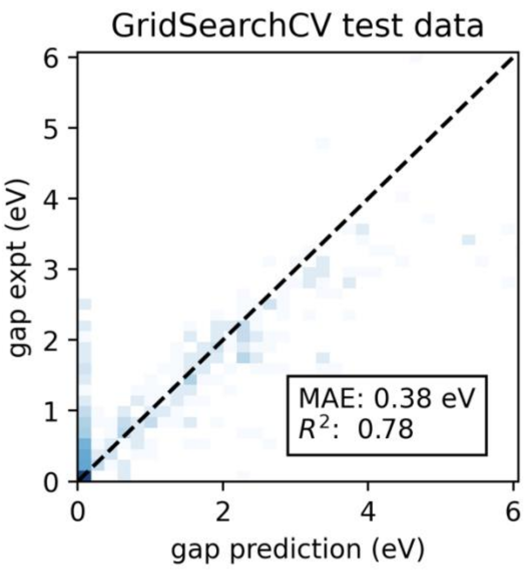


Figure 3. Prediction accuracy of the random forest regression model on test data from featurized matminer experimental band gap dataset.