Abstract

The goal of this project is to propose developing an open source model for predicting house prices in the United States. House price prediction has been a topic that Data Scientists have been discussing. The datasets that are from Kaggle, UCI Machine learning Repository. The datasets conclude features of the house, geographical attributes, etc. The main goal of the project is to create an accessible and accurate predictive model that can offer valuable insights to the US housing markets. Various regression modeling techniques such as Random Forest classifier will be used to generate accurate and effective predictive models, The project also emphasizes the significance of the quality of the dataset. Ensuring the predictive model is not only accurate but also accountable and unbiased. Analysis will be done at three distinct geographic levels: state, county, and city. A major focus will be the geospatial aspect of U.S. housing markets. This method improves the model's predictions and gives insight on the regional variables affecting home values. The project's open-source model offers cooperation, enabling developers, companies, and real estate agents to all work together to contribute for the model's development. This open-source methodology seeks to empower a broad user base by offering a transparent and adaptable tool for housing price prediction. To ensure the project's relevance and influence in the ever-changing U.S. real estate sector, it will be essential to enable the continuous integration for the live information, innovate modeling methodologies, and adhere to ethical principles.

Introduction

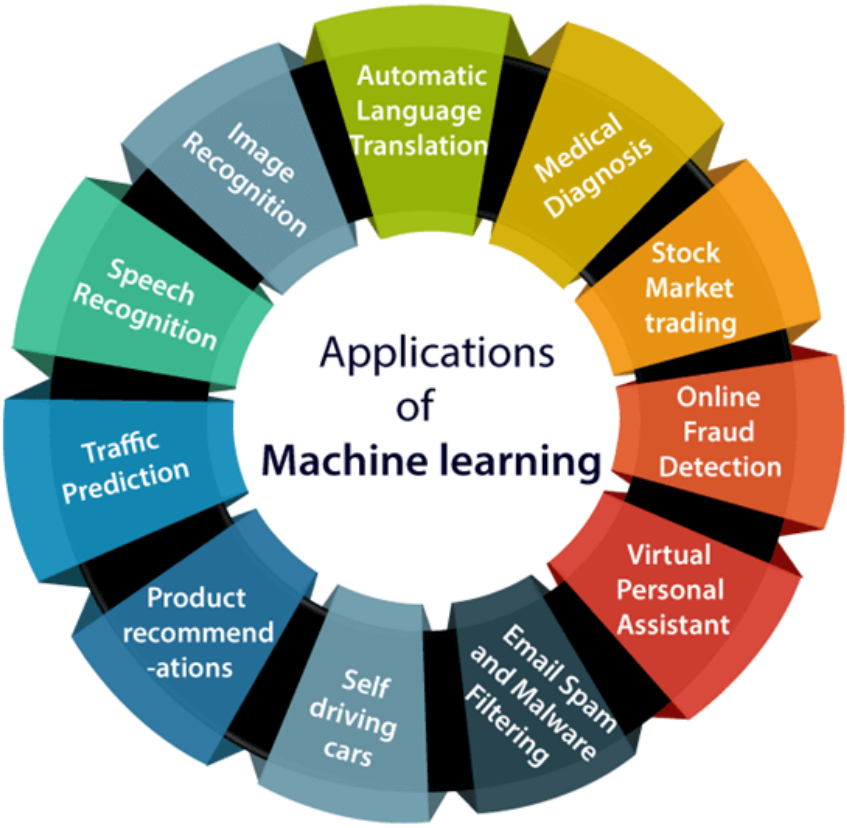
While researching the house prices in the United States because of the parent chores, My parent seem to struggle to decide whether it is a good time to buy or sell the house. How can house price prediction machine learning model can help people as an open source platform Informed Buying and Selling Decisions: Potential homebuyers can use house price predictions to make informed decisions about when and where to buy a property. Sellers can also benefit by pricing their homes competitively based on market trends. Financing: House price predictions can aid individuals in financial planning. Knowing the potential future value of a property can help homeowners plan for future expenses, investmenial Plats, and retirement. Real Estate Investment: Investors can use house price predictions to identify lucrative investment opportunities. Predicting the appreciation of property values helps investors make strategic decisions about buying, holding, or selling real estate assets. Risk Management: Financial institutions and mortgage lenders can use house price predictions to assess the risk associated with loans. This helps in making informed decisions about lending and managing potential financial risks

Background Research

Machine Learning is a branch of AI and Computer Science. By feeding information like us humans to the algorithm by use of the data, it improves the accuracy.
According to IBM “Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or

predictions, and to uncover key insights in data mining projects. These insights subsequently drive decision making within applications and businesses…."
There are various machine Learning techniques to build a model, but there are three major techniques that are commonly used. These techniques are based on statistical knowledge. First is Linear Regression, Linear regression is a technique that is used to study the relationship between independent variables and target variables. next is Logistic Regression, this is also the regression and lastly it is Random Forest.

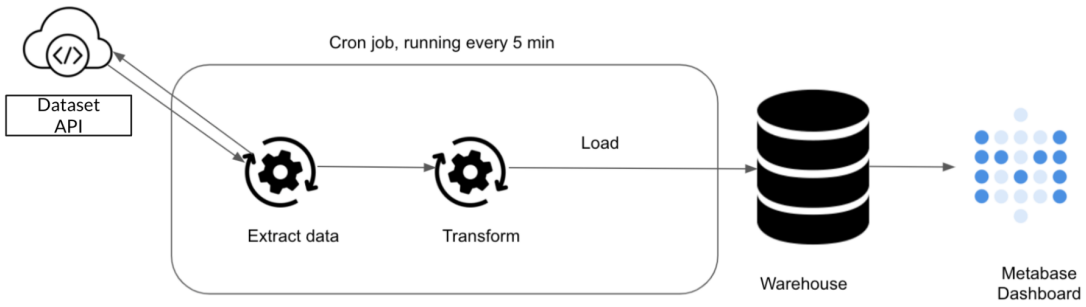
Below picture is the large application of Machine Learning. As we can see, a large application of machine learning is affecting all the industries and affecting or maybe in bad words threatening people's jobs. In many ways, they are more accurate than humans and do it cheaper than hiring a human. Machine learning's transformative impact on industries is undeniable, but it also raises concerns about its potential to automate certain tasks, potentially affecting employment. While ML systems offer increased accuracy and efficiency, there is an ongoing debate about the implications for the workforce and the need for reskilling in response to changing job requirements..



[Applications of Machine Learning - Javatpoint]

Data Dashboard

Data Dashboard seems to be the most effective way to have interactive communication with the general public who wants to buy or sell the house. Since, the main audience of the open source platform is for any users. This is the overall technical side of how the dashboard would work. Get the API from the dataset from the government house price prediction. extract the data using the data pipeline and transform these into the visualization and load it into the cloud platform, During this process we will have to manage the Continuous integration part to make the dataset live. After loading these databases into the warehouse, It will go to the metabase dashboard. Continuous integration may be a term that is unfamiliar, it is a code that extracts the dataset every 5 minutes or time we set and keeps it live.



[Designing a Data Project to Impress Hiring Managers · Start Data Engineering]

Target Audience

Target audience is mainly for the public because it is an open source platform, it will be a user -friendly interface that can be accessible to anyone and this will provide to the audience better understanding about the house price. Moreover, more specifically, it can also be for real estate professionals or consulting professionals. By having accurate predictions they can come up with an effective price strategy, asset value and market trends. Homebuyers and Sellers: Individuals looking to buy or sell properties can use the model to estimate fair market values, aiding in negotiations and informed decision-making. Investors: Real estate investors seeking opportunities in the U.S. market can leverage the predictive model to identify potentially lucrative areas and optimize their investment portfolios. Financial Institutions: Banks and mortgage lenders may find value in the model for risk assessment, loan underwriting, and determining property values in different regions. Policy Makers: Government agencies and policymakers can use the model to analyze housing market trends, inform housing policies, and address affordability and accessibility concerns

Goal

My vision for the project is to build an accurate and predictive model and to apply this to an open source model so anyone can access it. This model can be a tool for anyone who struggles to decide to buy or sell the house. Not only that by being an open source model, users can share the local information that only people would know by reporting and sharing. Furthermore, it can help real estate agents and consulting companies to convince the clients what it comes to purchasing or selling the house. By showing this, it can help prevent fraud by

having validation from agents and companies. My ultimate goal is also to build a community around a model that I built and improve and grow.

Development

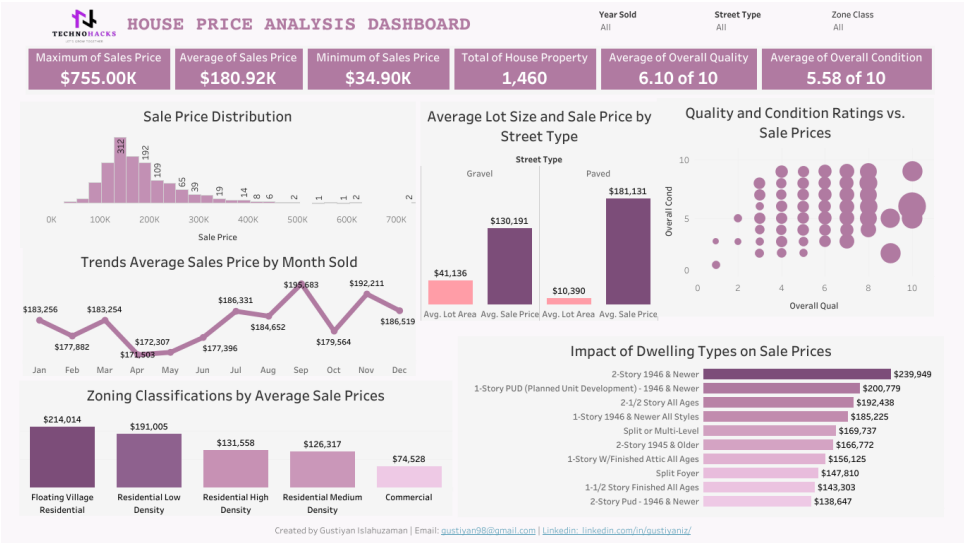
The community that first needs to be reached is the government. By gathering the dataset of the housing price from the government nationwide would have an accurate and legit machine learning model. Also collaborating with real-estate and consulting companies. It can help grow the model faster and have trust from the users. To do that Sharing a development process through my github repository. This can help us to build a prototype and get feedback during the pre-launch stage and even after the launch stage. Getting constructive feedback from the users matters for all. Feedback such as request form and common questions Q&A would be helpful.

Building a sustainable machine learning model is a first priority using python. I have built a model using a Random forest with a tensorflow and it seems like it has good RMSE for the dataset. Data preprocessing also needs to be done before building a model as well. Furthermore, not only random forest but XG/Gradient boosting can be used and compare the model RMSE/MSE. This needs to be further developed based on the dataset we get. In order to develop a nearly perfect predictive model, various developers will work collaboratively.

```python
import numpy as np

evaluation = rf.evaluate(x=valid_ds, return_dict=True)

# Print metrics including RMSE if available
for name, value in evaluation.items():
    print(f"{name}: {value:.4f}")

# Calculate and print RMSE if MSE is available
if 'mse' in evaluation:
    mse = evaluation['mse']
    rmse = np.sqrt(mse)
    print(f"RMSE: {rmse:.4f}")
else:
    print("MSE not found in evaluation results.")

```

```
1/1 [==============================] - 0s 144ms/step - loss: 0.0000e+00 - mse: 661278656.0000
loss: 0.0000
mse: 661278656.0000
RMSE: 25715.3389
```

To make it readable to all the users, visualizing all the gathered data would be the best. Most effective way is a Dashboard, where we build a data pipeline and visualize the statistics as you can see below. The tool is called Metabase which is an open source tool that allows for powerful data instrumentation, visualization, and querying. In order to update the datasets that we are receiving, we also have to set up continuous integration to the data pipeline via one of the cloud services, which is AWS. By setting up the CI, everytime we update the data, it will immediately apply to the pipeline



Security and privacy

Security and privacy should be tackled when it comes to working with a government. Having an open source model and working with a government may be tough but by talking whathere which data should be open to the public, it won't lead to massive data leak from the platform. By appropriately managing the permission by separating root users, which is having all the controls of the system for exclusive developers and IAM users for people like junior developers will be important. Users will have less amount of controls compared to other roles.

Project Expansion

This is the possible project expansion. If this project goes successfully, it can go up further by working with banks, consulting companies, Real-Estate agents. People who migrate from other countries may have trouble finding a house. With an integration of the bank account. If the family has money to buy the house but don't know where and how they can buy the house. My Open source platform can hook them up with a bank, consulting companies, and a real estate agent and show them a dashboard that it was made. If the family has trouble communicating, We can partner with a translating company and tag them along for a better service.

Conclusion

In conclusion, the development of an open-source model for predicting house prices in the United States emerges as a pivotal initiative with far-reaching implications. The inspiration for this project stems from a recognition of the challenges faced by individuals, including my own parents, in making informed decisions about buying or selling a home. By harnessing the power of machine learning and data science, this endeavor aims to provide an accessible and accurate predictive tool that empowers a diverse user base. The significance of datasets sourced from platforms like Kaggle and the UCI Machine Learning Repository cannot be overstated. These datasets, constraints various features of houses and geographical attributes, serve as the backbone of the predictive model. The choice of regression modeling techniques, particularly the utilization of Random Forest classifiers, underscores a commitment to generating accurate and effective predictions. It is acknowledged that the quality of the dataset is paramount, and the project places a strong emphasis on ensuring the resultant predictive

model is not only accurate but also accountable and unbiased. The geographic analysis at the state, county, and city levels enriches the model's understanding of regional dynamics, contributing to more nuanced and precise predictions. The geospatial aspect is not merely a technical detail but a strategic focus, enhancing the model's predictive capabilities by capturing localized factors influencing home values. The open-source nature of the project invites collaboration and contributions from a diverse community. Real estate professionals, investors, financial institutions, policymakers, and the general public constitute the primary audience. The model's transparency, accessibility, and adaptability make it a valuable tool for various stakeholders in the real estate sector. Beyond the technical aspects, the project envisions a broader impact on decision-making processes related to home buying and selling. The model has the potential to facilitate informed financial planning, assist real estate investors in identifying opportunities, and contribute to risk assessment for financial institutions. Security and privacy considerations, especially when collaborating with government datasets, are acknowledged and addressed through appropriate permissions and access controls. Looking ahead, the project's expansion envisions partnerships with banks, consulting companies, and real estate agents, aiming to assist individuals navigating the complexities of relocating and purchasing homes. In essence, this open-source predictive model is not just about predicting house prices; it represents an effort to democratize access to crucial information, foster collaboration, and empower individuals and professionals in making well-informed decisions in the dynamic landscape of the U.S. real estate market.

References

2uadmin. “What Is Machine Learning (ML)?” *UCB-UMT*, 19 Apr. 2022, ischoolonline.berkeley.edu/blog/what-is-machine-learning/.

“Applications of Machine Learning - Javatpoint.” *Www.Javatpoint.Com*, www.javatpoint.com/applications-of-machine-learning. Accessed 5 Dec. 2023.

“What Is Machine Learning?” *IBM*, www.ibm.com/topics/machine-learning. Accessed 5 Dec. 2023.