UNC CHARLOTTE
College of Computing and Informatics

Housing Price Prediction

A Case Study In Mecklenburg County

MIDTERM REPORT

Group Members

NAME	MAJOR	ID
Providence Adu	PhD Geography	padu@uncc.edu
Prashanth Minkuri	Master of Computer Science	pminkuri@uncc.edu
Xi Ning	PhD Applied Mathematics	xning1@uncc.edu

Problems and Challenges

- Introduction: Housing market is always an important topic in US as it reflects current economic situation as well as social sentiments. Also, players in housing markets often rely on averages to estimate housing prices. Hence, Having a model that predicts housing prices accurately can help both private individuals and local governments to have reliable estimates for housing prices.

- The Problem: Housing prices are impacted by multiple factors including location, ability to appreciate, age of the house, and square footage. This makes housing prices harder to predict as relationships among price impact variables can be non-linear. In addition, no work reports the use of machine learning (ML) algorithms to predict the values of properties in the Charlotte, Mecklenburg County.

- Challenges: It's hard to effectively capture the multi-collinearity and non-linearity among the variables. Sometimes, it also involves statistical assumptions in the samples. Therefore, the results of forecasting the housing prices generated might be unfavorable.

- How we solve: We intend to create a housing price prediction model that handles linear and non-linear relation between housing prices and also captures local variations to accurately predict housing prices in Charlotte, Mecklenburg County.

Approaches

- The goal of this project is to create a robust housing price prediction model for Charlotte, Mecklenburg County. Towards this end, we have mined data from the Mecklenburg County open data portal. The dataset has historic and current housing sales data in the county. The dataset has variables like number of bedrooms, date of sale, price, and number of stories.

- With this dataset, we have developed three (temporal, spatial, and variable combination) main preprocessing strategies in the creation of our price prediction model. These strategies were informed by our literature review and team brainstorming.

- We are preprocessing our data by categorizing our data based on geography and time. We call this **spatial** and **temporal** preprocessing. This categorization is based on the fact that Charlotte is a unique city hence when creating such a model for the city, we have to capture such uniqueness. This uniqueness comes from the fact that Charlotte's geographic patterns have shown that there is a divide in the city in terms of race and income. This has been popularly known as the wedge and the crescent. Datasets will be split into two main categories based on the spatial characteristics of Charlotte.

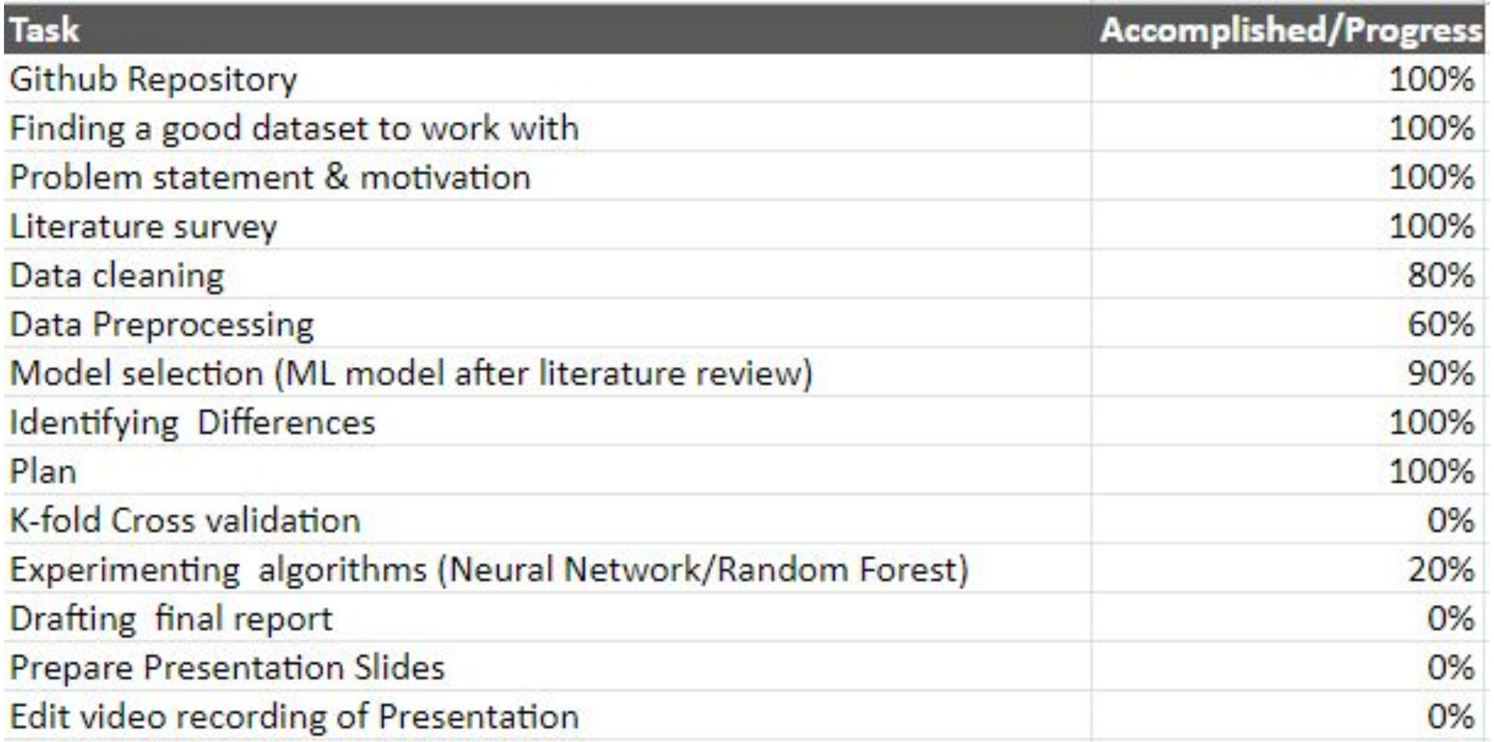
- In terms of the temporal preprocessing, we are looking at housing price variation relative to the US housing market crash in 2008. We want to assess our model estimates with respect to this housing crash time frame. Here we randomly select samples from sales pre 2008 and sales post-2008 for training and testing.

Approaches Cont’d

- The last strategy is consolidation of variables in our datasets into one single variable considering the fact that some housing characteristics correlated. For example the number of bedrooms, acres and square footage. We use variance inflation factor (VIF) to detect multicollinearity and lasso regression to select features since there are 70 features.

- Our literature review has shown that Random forest and Neural networks perform best when undertaking housing price prediction. Once we are done with preprocessing step, we are testing the performance using these two machine learning models. In addition, we will use K-fold cross validation to identify the best parameters for our models and we will train our models on our three different preprocessed data. Afterwards, we will undertake model comparison in terms of performance, and select the best model for our price prediction.

Accomplished and Incomplete Milestones



Task	Accomplished/Progress
Github Repository	100%
Finding a good dataset to work with	100%
Problem statement & motivation	100%
Literature survey	100%
Data cleaning	80%
Data Preprocessing	60%
Model selection (ML model after literature review)	90%
Identifying Differences	100%
Plan	100%
K-fold Cross validation	0%
Experimenting algorithms (Neural Network/Random Forest)	20%
Drafting final report	0%
Prepare Presentation Slides	0%
Edit video recording of Presentation	0%

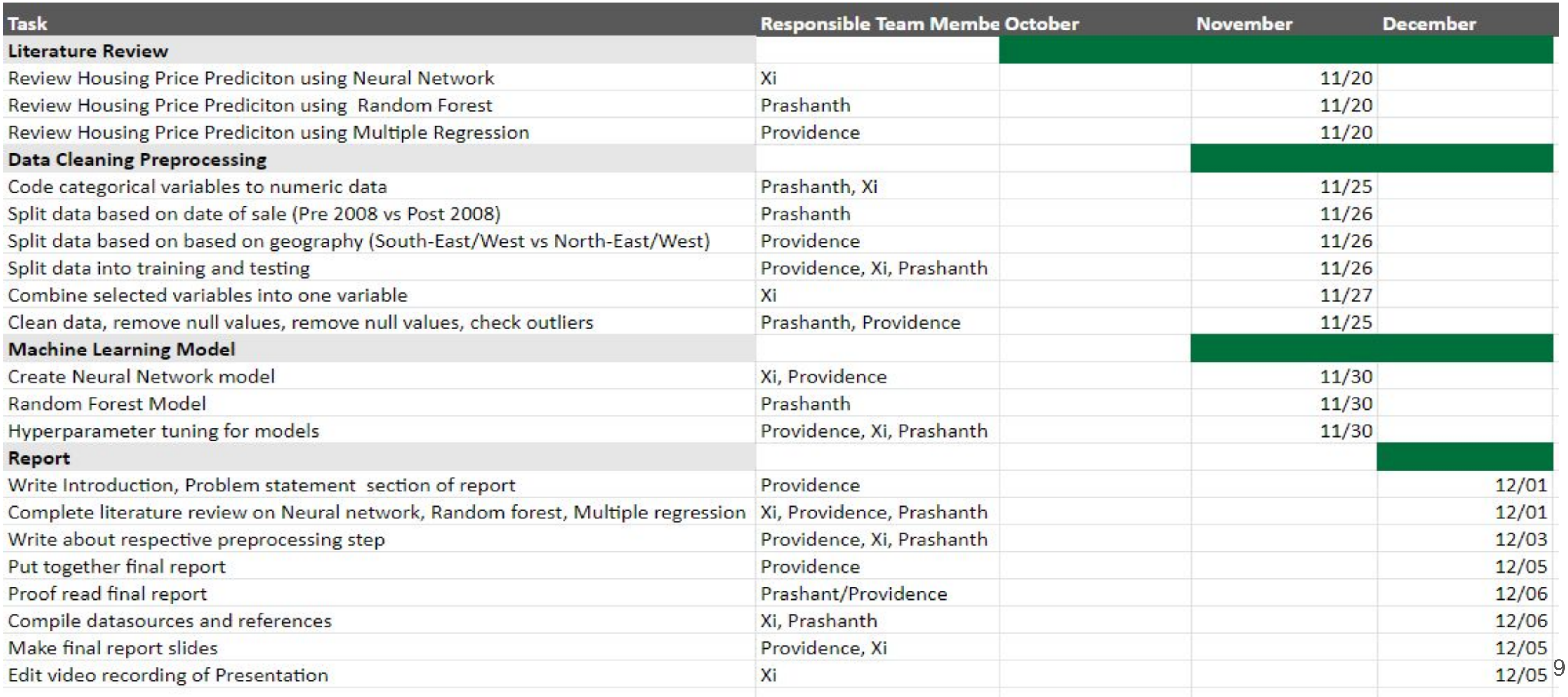
Difficulties

- There are more than 70 variables in our dataset and multiple categorical fields in the raw dataset that needs recoding

- Picking the best Machine Learning algorithm and determining structure of neural network

- Identifying the variables to combine and final input variables

Updated Plan



Task	Responsible Team Membe	October	November	December
Literature Review		🟩	🟩	🟩
Review Housing Price Prediciton using Neural Network	Xi		11/20	
Review Housing Price Prediciton using Random Forest	Prashanth		11/20	
Review Housing Price Prediciton using Multiple Regression	Providence		11/20	
Data Cleaning Preprocessing			🟩	🟩
Code categorical variables to numeric data	Prashanth, Xi		11/25	
Split data based on date of sale (Pre 2008 vs Post 2008)	Prashanth		11/26	
Split data based on based on geography (South-East/West vs North-East/West)	Providence		11/26	
Split data into training and testing	Providence, Xi, Prashanth		11/26	
Combine selected variables into one variable	Xi		11/27	
Clean data, remove null values, remove null values, check outliers	Prashanth, Providence		11/25	
Machine Learning Model			🟩	🟩
Create Neural Network model	Xi, Providence		11/30	
Random Forest Model	Prashanth		11/30	
Hyperparameter tuning for models	Providence, Xi, Prashanth		11/30	
Report				🟩
Write Introduction, Problem statement section of report	Providence			12/01
Complete literature review on Neural network, Random forest, Multiple regression	Xi, Providence, Prashanth			12/01
Write about respective preprocessing step	Providence, Xi, Prashanth			12/03
Put together final report	Providence			12/05
Proof read final report	Prashant/Providence			12/06
Compile datasources and references	Xi, Prashanth			12/06
Make final report slides	Providence, Xi			12/05
Edit video recording of Presentation	Xi			12/05

Feedback

Comment 1: Any plan for dealing with this issue? (referring to the Problem Statement where we mentioned multiple factors for housing prices)

- Undertake feature selection to see which variables are the most predictive in our study area. We lasso regression and Variance inflation factor to identify the most predictive variables in our datasets.

- Undertake literature review to see variables that are consistent throughout housing price prediction research.

Feedback

Comment 2: Provide link or source to data?

- Resolved by inserting a link
- The data has about 410,405 rows and 78 columns
  - Data source link (Mecklenburg County Open data Portal) - October 5th 2020
  - Data Dictionary In GitHub Repository
  - Raw Data On Google Drive

Feedback

Comment 3 : What do you think of their conclusion? (referring to the Literature Review we mentioned in project proposal)

- The MLR is much easier to implement but ANN has the ability to learn and model non-linear and complex relationships.

- In contrast to the MLR, ANN lacks explainability of a model no matter how accurate the final outputs are, which is known as a black-box issue.

- In the real world, we should preprocess the data to see if there are any patterns and choose a model based on your data.

Feedback

Comment 4: What is your overall synthesized summary? (referring to the Pros and Related Approaches in project proposal)

- If the dataset has many dimensions and medium size dataset SVM is good choice.
- If our objective is more on performance Neural Networks and Random Forest are best models. They perform really good on large data set.
- Random Forests provide highest accuracy rate however they take lot of memory for large datasets.
- Logistic Regression is simple to implement but they are not good model for non-linear data.
- If the focus of our model is on optimizing and reducing loss function Gradient Boosting Regression provides good hyperparameter tuning options.

Feedback

Comment 5: What is your direction given these cons? (referring to the Cons and Related Approaches in project proposal)

- From our literature survey we have found that advantages of Neural Networks are that they perform better when large data sample size is provided.
- Given our dataset has large number of records (410,405 rows) we have decided to apply Neural Networks as of one our models.
- Similarly, we have studied Random Forest classifiers provides good accuracy we will use this model on our dataset.

Feedback

Comment 6: What is your method? This is missing here. (referring to Methods in project proposal)

- Resolved by adding the detailed Approach in the midterm report (please see slide No.5 and No.6).
- We are building a regression model for predicting housing prices in Charlotte, Mecklenburg County using random forest and neural network. We building this model using three preprocessing strategies that capture the variations in our study area while making sure we have a robust prediction model.

Feedback

Comment 7: Comparison cannot be the major task of a project although it can be a part. (referring to Research Questions in project proposal)

- Resolved by focusing on the three main preprocessing strategies instead of the comparison of the algorithms. And we will also compare the performance of ANN and Random Forest.

Feedback

Comment 8: Can you come up with different approach for this? (referring to expectations of what we will learn from this project)

- How to pre-process data for machine learning model by incorporating local variations.
- How are the housing prices impacted during the 2008 recession period and what geographical areas of Charlotte city have been impacted most during this period?
- How to pre-process data for machine learning model by incorporating temporal variations.

Feedback

Comment 9: For this plan, you need to specify subtasks to accomplish your projects not the submission report subsections.

- Resolved by mentioning a detailed plan with specified subtasks and deadlines.

Feedback

Comment 10: All these are not proposing any “difference” for the project. (referring to how our approach is different)

- Resolved by adding the detailed differences in the midterm report (please see slide No.20, No. 21 and No.22).

Differences

Difference 1 : Spatial

- We have developed three main preprocessing strategies. The first strategy is that datasets will be split into two main categories based on the spatial characteristics of Charlotte. Charlotte's geographic patterns have shown that there is a divide in the city in terms of race and income. This has been popularly known as the wedge and the crescent. The wedge represents the southern part of Charlotte which is predominantly white residents and high income individuals. Conversely, the wedge represents the northern, west and eastern part of the city where the majority of the residents are predominantly African Americans and low income. This divide makes Charlotte unique compared to other cities. Testing the robustness of the housing price model by sampling housing prices from these two different geographical areas is important. First, it would help us give a representative model that does not only predict housing prices accurately but also reflect the local variations in our area of study.

Differences Cont’d

Difference 2 : Temporal

- The second significant difference we are focusing on is the Temporal Preprocessing of our dataset. We will split our dataset in such a way that we will analyse how sales and prices in Charlotte area have changed before and after **2008 year when Great Recession** hit United States.

- We analyse how the housing prices have impacted due to Recession and study what geographical areas of Charlotte city have affected most.

Differences Cont’d

Difference 3 : Variables Combination

- Lastly, considering the fact that there are more than 70 features in our data set and there is multicollinearity among some housing characteristics in our dataset, such as the number of bedrooms, the number of bathrooms and square footage, we combine some features into a new feature and drop off these original features. We use variance inflation factor (VIF) to detect multicollinearity.

- We will create *location_score* and *status_score* two features , where the location_score is to measure how well the location of the house is and the *status_score* is to measure how well/status the house is.

References

- [1] Nghiep Nguyen and Al Cripps. (2001). Predicting Housing Value: A comparison of Multiple Regression Analysis and Artificial Neural Networks. Journal of Real Estate Research. Vol. 22, 313-336.

- [2] Park, Byeonghwa, and Jae Kwon Bae. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications 42.6: 2928-2934.

- [3] Ng, Aaron, and Marc Deisenroth. (2015). Machine learning for a London housing price prediction mobile application. Imperial College London.

- [4] Fernandez-Duran, Laura & Llorca, Alicia & Ruiz, N & Valero, S. & Botti, V.. (2011). The impact of location on housing prices: applying the Artificial Neural Network Model as an analytical tool. ERSA conference papers.

References

- [5] Núñez Tabales, Julia M.; Caridad y Ocerin, José María; Rey Carmona, Francisco J. (2013). Artificial neural networks for predicting real estate prices, Revista de Métodos Cuantitativos para la Economía y la Empresa, ISSN 1886-516X, Universidad Pablo de Olavide, Sevilla, Vol. 15, pp. 29-44

- [6] Azme Bin Khamis and Nur Khalidah Khalilah Binti Kamarudin. (2014). Comparative Study On Estimate House Price Using Statistical And Neural Network Model. International Journal of Scientific & Technology Research Volume 3, Issue 12:126-131.

- [7] Wang, L., Chan, F. F. , Wang, Y & Chang, Q. (2016). Predicting Public Housing Prices Using Delayed Neural Networks. Proceedings of the International Conference. 3589-3592.

References

- [8] Banerjee, D., Dutta, S. (2017). Predicting the House Price Direction Using Machine Learning Techniques. IEEE International Conference on Power, Control, Signals and Instrumentation Engineering.
- [9] Afonso et al. (2019). Housing Prices Prediction with a Deep Learning and Random Forest Ensemble.
- [10] Truong, Q., Nguyen, M., Dang, H., Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. Procedia Computer Science. 174, 433-442.

PROPOSAL SLIDES

(The next slides are the project proposal slides)

UNC CHARLOTTE
College of Computing and Informatics

Housing Price Prediction

A Case Study In Mecklenburg County

Group Members

NAME	MAJOR	ID
Providence Adu	PhD Geography	padu@uncc.edu
Prashanth Minkuri	Master of Computer Science	pminkuri@uncc.edu
Xi Ning	PhD Applied Mathematics	xning1@uncc.edu

PROBLEM STATEMENT AND MOTIVATION

Problem Statement

- Housing market is always an important topic in US as it reflects current economic situation as well as social sentiments

- Housing prices are impacted by multiple factors including location, ability to appreciate, age of the house, number of bedrooms, and square footage

- This makes housing prices harder to predict as relationships among price impact variables can be non-linear sometimes

- The non-linear nature of housing price variables make it hard to predict the prices houses, hence many models rely on housing price averages

- Traditional statistical approaches such as hedonic models don’t accurately predict housing prices, hence there is a lot of inaccuracies in predicted prices

Motivation

- Having a model that predict housing prices accurately can help both private individuals and local governments to have reliable estimates for housing pricing

- Having a model that aids in accurate housing price prediction can help ensure proactive decision making that addresses both economic and social needs

- Applying Machine Learning model to housing price prediction problem can help give insights to behavior of housing markets as well as how well different Machine learning models can be applied in solving social and economic problems

- Insights about the performance of Machine Learning approaches as compared to traditional statistical approaches can be gained from this analysis

Data

- The data for this research is tax parcel data from Mecklenburg County. The city government maintains a database of all the parcels in the Charlotte area

- The tax parcel data has attributes or features that describe: sales prices of houses on each parcel, number of bedrooms, owner of parcels, building grade, among other attributes.

- The data has about 410405 rows and 78 columns.

Literature Review

Summary of Related Approaches

Nguyen and Cripps (2001) – Compared Artificial Neural Network (ANN) vs. Multiple regression analysis

- They found that if one provides sufficient data training size and appropriate ANN parameters, then ANN performs better than MRA.
- For practical purposes, the ANN is recommended when there is sufficient sample data set and/or when there is no theoretical basis for the data model functional form. Otherwise, the MRA is recommended.

Summary of Related Approaches

Banerjee and Dutta (2017) – Compared Random Forest, Artificial Neural Network and Support Vector Machine

- The authors only performed a basic experiment comparing four metrics, accuracy precision sensitivity and specificity between the three models. However, in their paper they did not discuss any theoretical explanations for why one model may be better than another.

- In addition, they concluded that random forest was the best model because it had the highest accuracy.

- However, this conclusion could be misguided as accuracy is not necessarily a great metric to judge models. For example, if a dataset is extremely imbalanced, accuracy would not be the proper statistic.

Summary of Related Approaches

Afonso et al (2019) – Combined Random Forest and Recurrent Neural Networks

- The most striking issue in this paper involved their dataset. The authors dealt with mixed data types including images and pictures. As a result, they were forced to mix and match many different models together to produce an output. This many moving parts could cause many potential errors.

- Another negative of the data was that there were many missing values for key attributes making it hard to perform regression. For some attributes such as number of floors , over 90% of data was missing. In addition, there were a number of attributes that contained extreme values that were removed to make visualization easier.

Summary of Related Approaches

Quang Truong et al (2020) – Compared the performance of Random Forest, XGBoost and LightGBM

- The Random Forest method has the lowest error on the training set but is prone to be overfitting but its time complexity is high since the dataset has to be fit multiple times.

- The XGBoost and LightGBM are decent methods when comparing accuracy, but their time complexities are the best, especially LightGBM.

Pros of Related Approaches

- ANN performs better than the MRA when a moderate to large data sample size is used.
- Gradient Boosting Regression provides lots of flexibility - can optimize on different loss functions and provides several hyperparameter tuning options.
- Random forest classifiers provide the highest accuracy among all classification models.
- Support Vector Machines is better when there are more number of dimensions.

Cons of Related Approaches

- SVM does not perform well when there is more noise in the data.
- Random forest classifiers take a lot of memory for very large data sets.
- Gradient Boosting Models can overemphasize outliers and cause overfitting to reduce errors.
- Artificial neural networks require processors with parallel processing power.

METHODS

Research Questions

- What features are most predictive of housing prices?
- What type of model is the most accurate in predicting housing prices?
- How can we finetune our model and improve our performance to achieve better results?

Expectations Of What We Will Learn From This Project

- How to pre-process the data before applying a model on it
- How to undertake feature selection using methods like lasso regression
- How to prepare and train a Neural Network
- Which algorithm ensures better accuracy in housing price prediction
- How to use machine learning to predict housing price

Anticipated Approach/Steps

- **Data Cleaning**

  Remove null/0’s for price

  Remove/transform text columns

- **Feature Selection:**

  Lasso, PCA

- **Model Selection** :

  Multiple Linear Regression, Random Forest, Neural Network

- **Prediction** :

  Train data and predict housing prices

- **Models Comparison :**

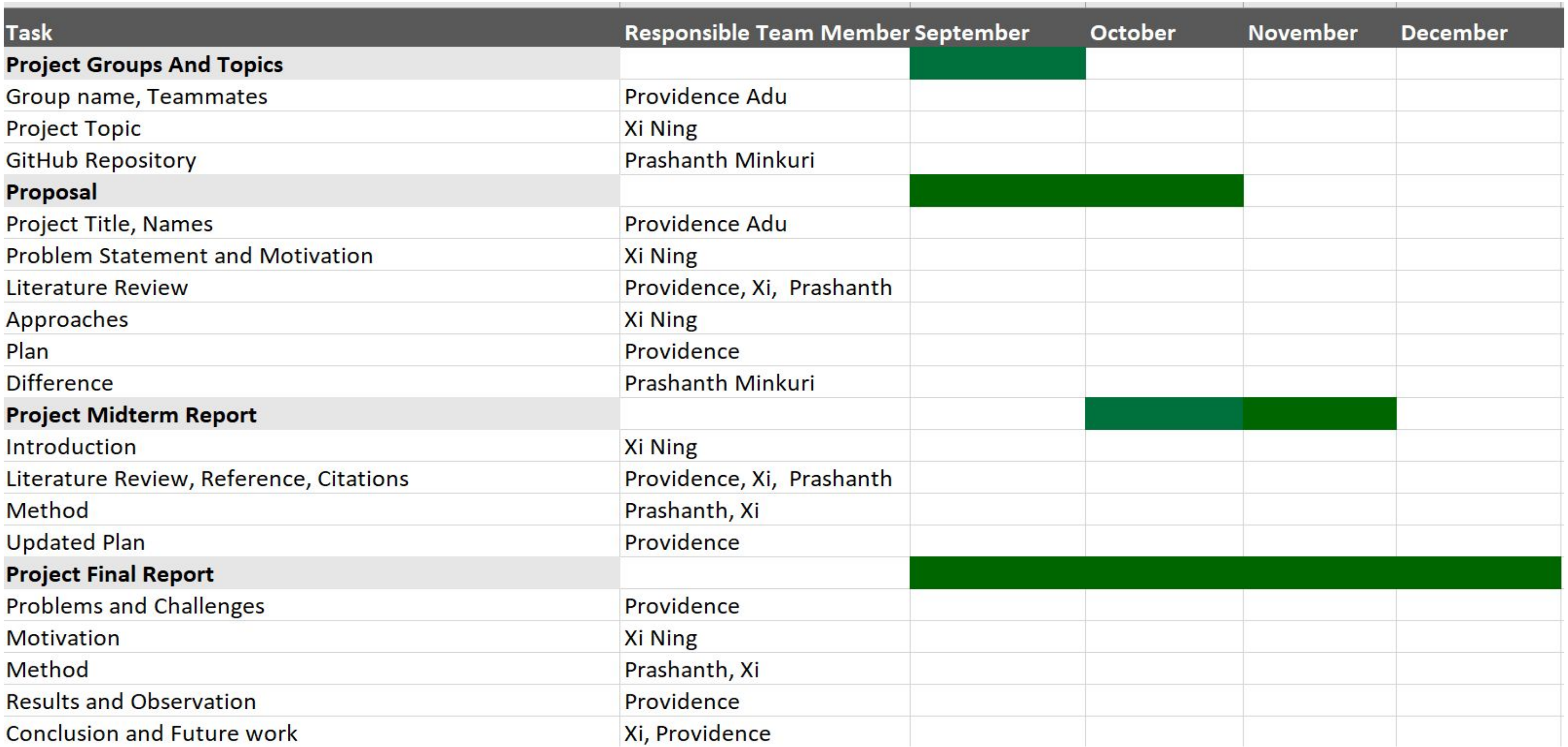
  Multiple Linear Regression vs Random Forest vs Neural Network

- **Comparison Approach:**

  K-fold cross-validation, Mean square error

PLAN

Plan



Task	Responsible Team Member	September	October	November	December
Project Groups And Topics		🟩			
Group name, Teammates	Providence Adu				
Project Topic	Xi Ning				
GitHub Repository	Prashanth Minkuri				
Proposal		🟩	🟩		
Project Title, Names	Providence Adu				
Problem Statement and Motivation	Xi Ning				
Literature Review	Providence, Xi, Prashanth				
Approaches	Xi Ning				
Plan	Providence				
Difference	Prashanth Minkuri				
Project Midterm Report			🟩	🟩	
Introduction	Xi Ning				
Literature Review, Reference, Citations	Providence, Xi, Prashanth				
Method	Prashanth, Xi				
Updated Plan	Providence				
Project Final Report		🟩	🟩	🟩	🟩
Problems and Challenges	Providence				
Motivation	Xi Ning				
Method	Prashanth, Xi				
Results and Observation	Providence				
Conclusion and Future work	Xi, Providence				

DIFFERENCE

How Our Approach Is Different

- **Location**: There are no research reports on the use of machine learning (ML) algorithms to predict the values of properties in the Charlotte area.

- **Variables and feature selection:** Our data set contains more than 400,000 observations and 70 features so feature selection is key. Unlike past approaches, we intend to use Lasso Regression and PCA to select the features. Also, we intend to include variables like income in our approach

- **Model comparison**: Unlike past approaches, we will use a more robust method (k-fold cross validation) to compare the three models – Multiple Linear Regression, Random Forest, and Neural Network.

- **Results Presentation**: We will create descriptive visualizations to describe results instead of only tables of statistics.

REFERENCES

References

- [1] Nghiep Nguyen and Al Cripps. (2001). Predicting Housing Value: A comparison of Multiple Regression Analysis and Artificial Neural Networks. Journal of Real Estate Research. Vol. 22, 313-336.
- [2] Park, Byeonghwa, and Jae Kwon Bae. "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data." Expert Systems with Applications 42.6 (2015): 2928-2934.
- [3] Ng, Aaron, and Marc Deisenroth. "Machine learning for a London housing price prediction mobile application." Imperial College London (2015).
- [4] Wang, L., Chan, F. F. , Wang, Y & Chang, Q. (2016). Predicting Public Housing Prices Using Delayed Neural Networks. Proceedings of the International Conference. 3589-3592.

References

- [5] Banerjee, D., Dutta, S. (2017). Predicting the House Price Direction Using Machine Learning Techniques. IEEE International Conference on Power, Control, Signals and Instrumentation Engineering.

- [6] Afonso et al. (2019). Housing Prices Prediction with a Deep Learning and Random Forest Ensemble.

- [7] Truong, Q., Nguyen, M., Dang, H., Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. Procedia Computer Science. 174, 433-442.