**Housing Price Prediction Project - Meeting Log**

**Date:** January 29, 2025 **Phase:** Project Proposal & Literature Survey

**Key Points**

* Selected housing price prediction project
* Decided to use real estate datasets
* Will compare multiple ML algorithms

**Discussion Summary**

Team settled on housing price prediction after discussion. Amarnath shared real estate industry experience. Will develop general model comparing various ML algorithms, starting with classical regression techniques as baselines.

**Data Sources**

* Primary focus on Zillow datasets
* Additional real estate data websites to be identified

**Next Steps**

* Research existing housing price prediction models
* Identify and collect datasets from Zillow and other sources
* Draft initial project proposal
* Set up project repository
* Team members to review literature in assigned areas

**Date:** February 14, 2025 **Phase:** Model Development & Baseline Implementation

**Key Points**

* RMSE and MAE as primary metrics
* Develop ensemble approach
* Use cloud computing resources

**Discussion Summary**

Baseline results: Linear Regression, Ridge, Decision Tree. Will implement KNN, SVR, Random Forest, Gradient Boosting, and Neural Networks. Location features dominate importance analysis.

**Model Performance (Baseline)**

* Linear Regression: RMSE evaluation completed
* Ridge Regression: RMSE evaluation completed
* Decision Tree: RMSE evaluation completed

**Evaluation Metrics**

* Primary: RMSE, MAE
* Secondary: MAPE (Mean Absolute Percentage Error)

**Next Steps**

* Implement and benchmark multiple regression models
* Set up cloud computing environment for model training
* Develop feature importance analysis framework
* Create visualization dashboard
* Prepare individual EDA reports
* Begin planning for machine learning model comparisons

**Date:** February 23, 2025 **Phase:** Algorithm Comparison & Feature Engineering

**Key Points**

* Gradient Boosting models best performing
* Feature engineering crucial
* Begin research paper

**Discussion Summary**

XGBoost and LightGBM outperformed other methods. Neighborhood aggregations and temporal trends reduced error by 15-20%. Location features most important (35% error increase when removed). XGBoost offers good balance of performance and interpretability.

**Model Performance**

* XGBoost: Best performance among models tested
* LightGBM: Very competitive performance
* Random Forest: Good but not best performance
* Linear Methods: Significantly higher error rates

**Feature Engineering Impact**

* Neighborhood aggregations and temporal trends: 15-20% error reduction
* Removing location features: 35% error increase
* Removing property characteristics: 28% error increase

**Next Steps**

* Conduct comprehensive hyperparameter tuning
* Draft research paper introduction and methodology
* Explore model explainability techniques like SHAP
* Implement advanced feature engineering
* Prepare for EDA presentations

**Date:** March 2, 2025 **Phase:** EDA Presentations (Group)

**Key Points**

* Add geographic visualization
* Improve model interpretability
* Maintain project timeline

**Discussion Summary**

Lochana suggested more geographic visualizations. Discussed mapping prediction errors as heatmaps. SHAP values helpful for personalized explanations. Time-series forecasting considered as stretch goal.

**Presentation Feedback**

* Positive feedback on thorough data exploration
* Recommendation to add more geographic visualizations
* Impressed with systematic approach to algorithm comparison

**Visualization Plans**

* Heatmaps of prediction errors by location
* SHAP value dashboards for personalized explanations
* Geographic visualization of price patterns and correlations

**Next Steps**

* Implement geographic visualizations for model performance
* Continue paper writing, focusing on results section
* Develop explanation dashboard using SHAP values
* Fine-tune model hyperparameters
* Prepare for AI/ML model evaluations
* Begin planning for time series modeling approaches

**Date:** March 15, 2025 **Phase:** Advanced AI/ML Models Presentations **Focus:** Machine Learning Models

**Key Points**

* Ensemble of gradient boosting models performs best
* Traditional ML models provide strong baselines
* Feature importance and model interpretability emphasized

**Discussion Summary**

Our comprehensive comparison of traditional machine learning models showed gradient boosting methods outperforming others. The ensemble approach combining multiple gradient boosting models achieved our best performance. Model interpretation using SHAP values was well-received by the group, allowing us to explain predictions in terms of feature contributions.

**Model Performance**

* Gradient Boosting Ensemble: Best performance
* Random Forest: Good performance
* Support Vector Regression: Moderate performance
* KNN: Lower performance

**Interpretability Approach**

* SHAP values for explaining individual predictions
* Feature importance analysis across different neighborhoods
* Interactive dashboard showing key factors for each prediction

**Next Steps**

* Fine-tune ensemble approach based on feedback
* Refine interpretation dashboard with additional visualizations
* Integrate with time series modeling component
* Prepare for LSTM implementation

**Date:** March 19, 2025 **Phase:** Advanced AI/ML Models Presentations **Focus:** Time Series Models

**Key Points**

* Time series models capture market dynamics
* ARIMA and Prophet models evaluated
* Temporal features improve prediction accuracy

**Discussion Summary**

We presented our analysis of housing market temporal dynamics using time series models. ARIMA models showed reasonable performance but struggled with spatial components. Facebook Prophet models handled seasonality well. Most importantly, we demonstrated that incorporating temporal features from these models into our main gradient boosting models improved overall performance.

**Model Performance**

* ARIMA: Moderate performance
* Prophet: Better handling of seasonality
* Gradient Boosting with temporal features: Best performance

**Key Time Series Features**

* Price momentum (3-month, 6-month, 12-month)
* Seasonality components
* Days-on-market trends
* Interest rate correlations

**Next Steps**

* Integrate time series predictions into ensemble approach
* Develop neighborhood-specific time series forecasts
* Incorporate economic indicators as exogenous variables
* Prepare for LSTM implementation
* Begin developing hybrid model architecture

**Date:** March 16, 2025 **Phase:** Advanced AI/ML Models Presentations **Focus:** LSTM Update & Mid-term Feedback

**Key Points**

* LSTM models show promise but require more data
* Addressing mid-term feedback on model evaluation
* Hybrid approach combining LSTM with gradient boosting

**Discussion Summary**

Vikas presented results from neural network experiments, focusing on LSTM architectures for capturing temporal dependencies in housing prices. While deep learning approaches performed competitively, they didn't outperform our ensemble of gradient boosting models despite requiring significantly more computational resources. We discussed strategies to address mid-term feedback regarding cross-validation and model generalization.

**Model Performance**

* LSTM Neural Network: Competitive performance
* Gradient Boosting Ensemble: Better performance
* Hybrid (LSTM features + Gradient Boosting): Best performance

**Mid-term Feedback Addressed**

* Implemented geospatial cross-validation
* Added confidence intervals to predictions
* Enhanced interpretability with attention mechanisms
* Improved documentation of model architecture

**Next Steps**

* Collect additional data to better leverage LSTM capabilities
* Refine hybrid model architecture
* Implement attention mechanisms for interpretability
* Address remaining mid-term feedback points
* Explore specialized models for different property types/price segments

**Date:** April 1, 2025 **Phase:** Data Ethics & Presentation Skills

**Key Points**

* Address potential model biases
* Enhance visualizations
* Submit drafted project report

**Discussion Summary**

Discussed ethical implications of housing price prediction, including historical biases. Model showed varying error rates across neighborhoods. Explored fair regression techniques and visualization improvements. Demonstrated prototype application.

**Ethical Considerations**

* Potential reinforcement of historical biases in real estate
* Varying error rates across different neighborhood demographics
* Need for balanced performance across all areas
* Transparency in model predictions

**Visualization Improvements**

* Clear geographic displays of predictions
* Avoid misleading charts and representations
* Proper confidence interval visualization
* Interactive exploration of feature importance

**Next Steps**

* Conduct fairness analysis across neighborhood demographics
* Refine presentation materials and visualizations
* Incorporate ethical considerations into drafted project report
* Implement fairness-aware model adjustments
* Submit drafted project report

**Date:** April 14, 2025 **Phase:** Project working week

**Key Points**

* System integration complete
* Final report structure agreed
* API design finalized

**Discussion Summary**

Fixed data pipeline issues and improved documentation. Report will emphasize technical innovation and practical applications. UI now includes confidence intervals and property recommendations. Model outperforms published research by 7-12%.

**System Improvements**

* Fixed data pipeline reproducibility issues
* Comprehensive documentation added
* Improved user interface with confidence intervals
* Added comparable property recommendations feature

**Report Structure**

* Technical innovation in feature engineering
* Practical applications for real estate professionals
* Model comparison methodology and results
* Ethical considerations and fairness analysis
* Future work and improvements

**Next Steps**

* Finalize benchmark comparisons against published research
* Draft final report structure and introduction
* Complete user interface for demonstration
* Implement confidence intervals for predictions
* Prepare for Data Science Competition

**Date:** April 23, 2025 **Phase:** Project working week

**Key Points**

* System fully tested
* Report draft completed
* Presentation materials prepared

**Discussion Summary**

Conducted end-to-end testing, fixed API and frontend issues. Feature engineering provided more improvement than algorithm selection. Presentation will highlight systematic model comparison and interpretability tools.

**Testing Results**

* End-to-end system testing completed
* Minor API and frontend issues fixed
* All data pipelines functioning correctly
* Model serving performance optimized

**Lessons Learned**

* Feature engineering provided more substantial improvements than algorithm selection
* Early focus on data quality pays significant dividends
* Geospatial cross-validation essential for real-world performance assessment
* Interpretability as important as raw performance metrics

**Next Steps**

* Complete system testing and documentation
* Finalize report draft
* Prepare presentation materials
* Set up GitHub repository with code and documentation
* Practice final presentation