GA DSI 26 Project 2: Ames Housing Prices

By: Lim Zhi Yong

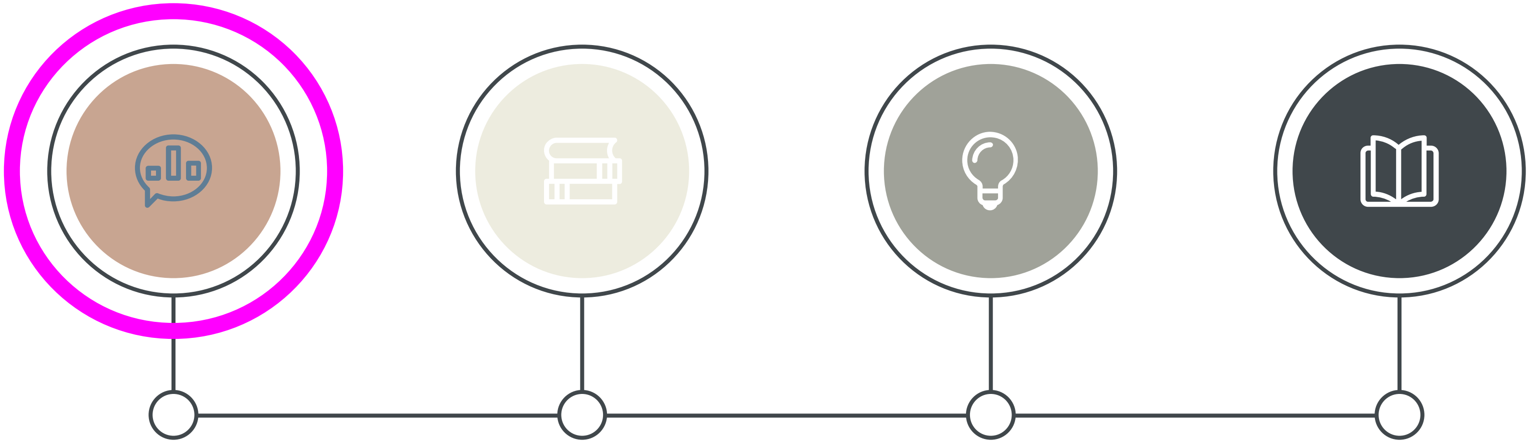
Task:

- Which features improve housing prices
- Which features negatively impact prices
- To build a model for the prediction of housing prices

Data Description

- 2051 rows, 81 columns (80 for test)
- Based in Ames, IA
- Data taken from 2006-2010

Model Workflow



Data Cleaning

- Missing values
- Creating new features
- Choosing features

Modelling

- Building models
- Scoring models

Testing

- Kaggle testing on test dataset

Recommendations

- Important features

Missing Values

There are different types of missing values:

- Impute 0 for no garage, basement etc
- KNN impute for things that are supposed to be there eg. lot frontage
- Drop features with too many missing values and too low correlation

Conscious effort not to drop rows unnecessarily (only 2 dropped)

Creating Features

Some features were created, e.g.

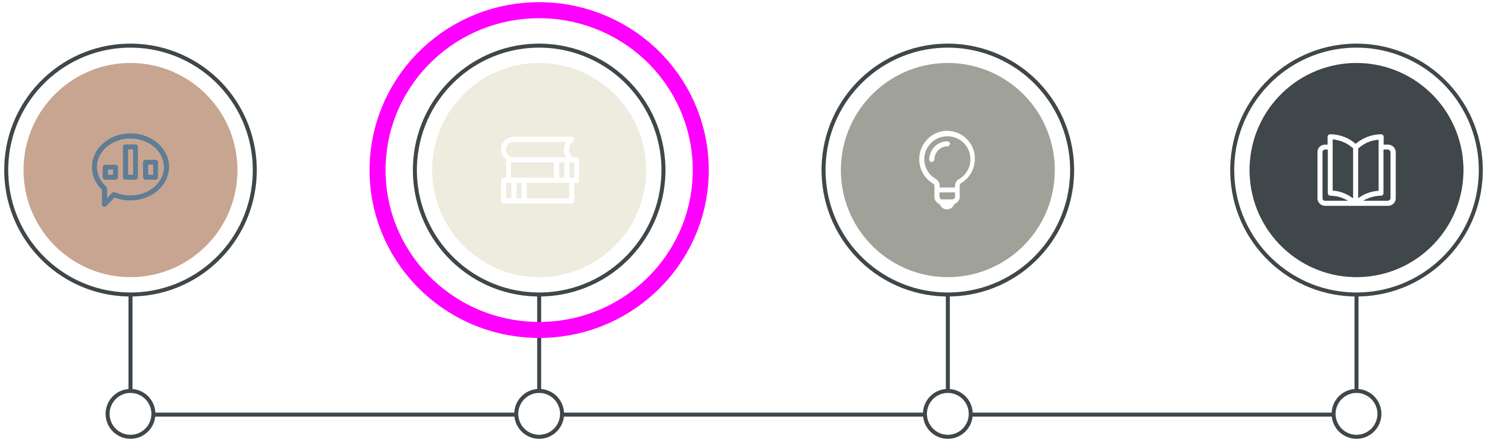
- Ages were calculated instead of using raw years
- Nominal variables were one-hot encoded
- Ordinal variables were ranked numerically

Choosing Features

The conditions for choosing were:

- Correlation of above 0.4 with sale price (some exceptions)
- Not directly related to other variables (independence)
- Normalize continuous/discrete data by removing outliers (winsorization or log transform)

Model Workflow



Data Cleaning

- Missing values
- Creating new features
- Choosing features

Modelling

- Building models
- Scoring models

Testing

- Kaggle testing on test dataset

Recommendations

- Important features

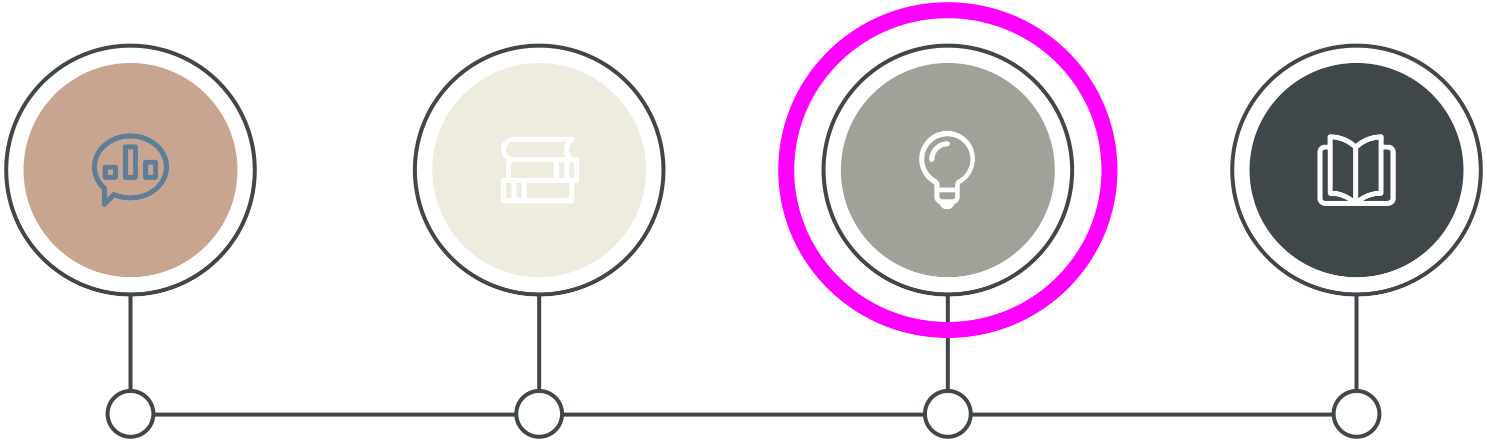
Modelling

- 4 models:
  - Linear regression
  - LASSO
  - Ridge
  - Elastic net

Scoring

- Linear regression model performed the best
- Scored with RMSE

Model Workflow



Data Cleaning

- Missing values
- Creating new features
- Choosing features

Modelling

- Building models
- Scoring models

Testing

- Kaggle testing on test dataset

Recommendations

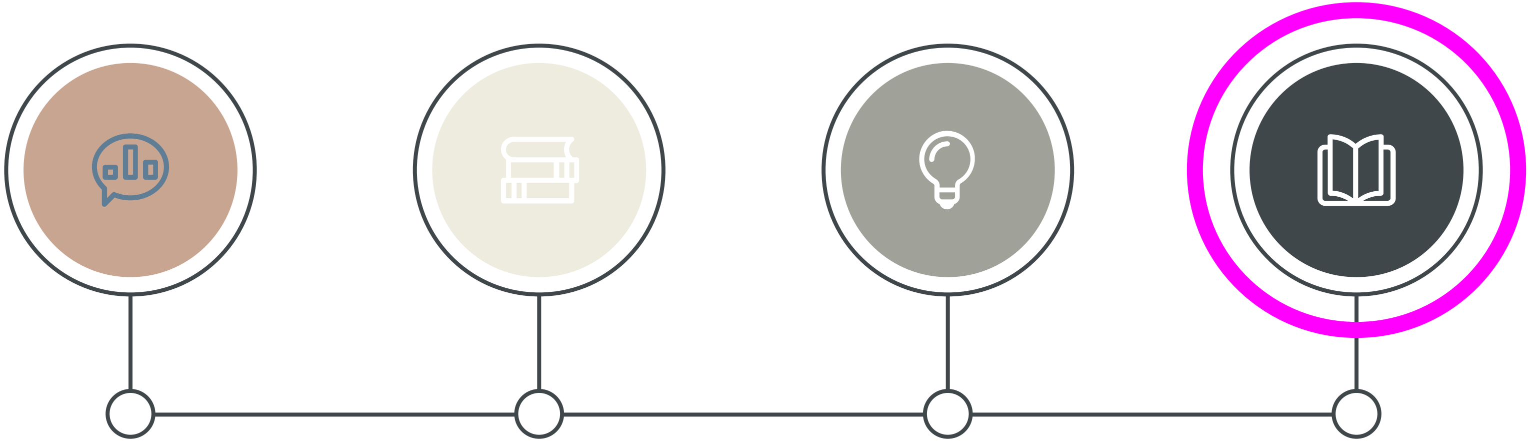
- Important features

Linear Regression performed the best

22,038

RMSE on Kaggle's leaderboard (rank: 28)

Model Workflow



Data Cleaning

- Missing values
- Creating new features
- Choosing features

Modelling

- Building models
- Scoring models

Testing

- Kaggle testing on test dataset

Recommendations

- Important features

Recommendations

Important features

- Exterior quality, material, masonry veneer, and finish
- Gross living area, together with garage and bedroom
- Amenities, e.g. central air-conditioning
- Lot area and frontage
- Neighborhood

Negative

- Age of house/garage: the older, the cheaper it is
- Roof, deck/porch, pool: low impact

Thanks

CREDITS: This presentation template was created by **Slidesgo,** including icons by **Flaticon,** infographics & images by **Freepik**

Please, keep this slide for the attribution