I. Definition

III. Methodology

V. Conclusion

II. Preparation

IV. Results

| Structuring a Machine Learning Project | | | |
| --- | --- | --- | --- |
|  | Domain  Liquor sales data in Iowa.  Of a specific subset of licensees though the state of Iowa, 2015 - Q1 2016. | Data Preprocessing   1. Verify the quality of the data - are there errors? missing data/data completeness? data mismatch? data over/under representation 2. Preprocess the Raw Data making calculated/derived fields (e.g. log transform, poly transform), dummy variables, casting to numerical types, reduce dimensionality 3. Feature Selection by Forward Selection, Backward Selection, Grid Search |  |
| Problem Statement  Predict 2016 sales by creating a regression model (e.g. OLS, SGD, Lasso, Ridge, ElasticNet) using Q1 2015 sales and the entire year to create a model for the entire year of 2016.  Predict the optimal opening location of a new store by creating a regression model (e.g. OLS, SGD, Lasso, Ridge, ElasticNet) using same sales data. Or by identifying prime location using all data. | Implementation  load the data from csv using pandas  Visualize using seaborn  Model and Score using sklearn |
| Metric  R2 (Coefficient of Determination), Mean Squared Error, Mean Absolute Error | Refinement  **Tried, not so good**  tried OLS which gave negative r squared  fit on overly correlated features (no new information)  added another dataset (unemployment data), added v little  Created dummy variables to include categorial data  Benchmark not clear (maybe some articles on the site)  **Worked**  Premium alcohol was a good feature  Identifying by type of alcohol  number of stores  pandas pivot tables worked well |
|  | Data Exploration  Read the README. What types of data (ints, floats, strings, booleans, etc)/ astype. What is necessary/extraneous? Pivot tables. groupby, agg, check for garbage data (GIGO), measures of central tendency, densities, describe(), info, unique, shape, recategorization, sort, | Model Evaluation and Validation  Rather than sales, look at changes in sales over time in order to identify where the greatest demand will be. 520xx  (delta sales/zip\_code/year\_i) -> lasso  use built model to predict on full dataset, looked for largest delta |  |
| Visualization  sns.barplot, sns.heatmap, sns.pairplot, pd.scatterplot, sns.countplot, plt.hist, plt.plot, sns.stripplot, sns.swarmplot, sns.violinplot | Justification |
| Algorithms and Techniques  Ordinary Least Squares, optionally apply regularization using l2norm (ridge), l1norm (lasso), both (elastic).  Stochastic Gradient Descent, optionally apply regularization  Pivot Tables and groupby-aggregate function, merge-join,  Heavy emphasis on data processing, large messy dataset | Reflection |  |
| Benchmark  Performance with respect to metric selected above. | Visualization |
| Code Design   1. Load the Data in Pandas 2. Clean/Munge the Data - get rid/impute nans, get rid of errant strings, 3. Verify the quality of the data - are there errors? missing data/data completeness? data mismatch? data over/under representation 4. Preprocess the Raw Data making calculated/derived fields (e.g. log transform, poly transform), dummy variables, casting to numerical types, reduce dimensionality 5. Feature Selection by Forward Selection, Backward Selection, Grid Search 6. Split the data into training, test sets 7. Normalization of the data 8. Create a new model 9. Fit the model 10. Score the model | Recommendations |