Predicting Real Estate Sales Using Machine Learning and Spatial Dependence

Contents

Introduction

Introduction to my thesis

Income inequality may be a central challenge of our time. Researchers at the Urban Institute recently identified the socio-economic phenomenon of "Economic Exclusion" as one compelling explanation for the recent rise in inequality in the US. Vulnerable populations (disproportionately communities of color, immigrants, refugees, and women), who are displaced by localized economic prosperity enter into a gradual cycle of diminished access to good jobs, good schools, health care facilities, public spaces, etc. This systematic denial causes enduring and self-reinforcing poverty over the course years and even generations (Greene, et al. 2016), gradually entrenching income inequality and general unrest.

One way to practically combat economic exclusion is to focus on preventing displacement, however, detecting gentrification at an early enough stage can be a daunting task. When an area experiences economic growth, increased housing demands and subsequent affordability pressures can lead to evictions of low-income families and small businesses. Government agencies and nonprofits tend to intervene once displacement is already underway, and after-the-fact interventions can be costly and ineffective. There are a host of pre-emptive actions that can be deployed to stem divestment and ensure that existing residents benefit from new investments. Not unlike medical treatment, early detection is key to success. Consequently, in 2016, the Urban Institute put forth a call for research into the creation of "neighborhood-level early warning and response systems that can help city leaders and community advocates get ahead of neighborhood changes."

This paper explores novel techniques to predict gentrification in the pursuit of combating displacement and economic exclusion. Modern techniques of data mining, machine learning and predictive modeling are applied to datasets describing property values and sale prices in New York City. I explore the viability of using spatial lags, i.e., variables created from physically proximate observations, as features in a machine learning predictive model.

Literature Review

Lit Review

The world has seen an unprecedented amount of geospatial data produced in recent years (i.e., data that contain information about where an observation exists or happened). Every day in the U.S., federal, state and local government agencies are making their troves of geo-spatially tagged data available for the benefit of the public. Adequate tools to describe, explore and model such data are in short supply for the data-journalists and data-activists who have become modern mechanisms of public service. It is imperative that research be done and tools created to better harness such data for commercial and public good.

Predictive modeling using spatial dependence has been employed extensively in recent years, notably in Crime Prediction (Almanie et al., 2015). However, a key deficiency of many spatial dependent models are their use of arbitrarily defined geographic regions, such as zip codes, political districts, police precincts, state lines, neighborhoods, etc. which diminish and obscure potentially valuable insights. Worse yet, many predictive models ignore and exclude spatial dependence, violating one of the basic tenets of geography: the direct relationship between distance and likeness (Miller, et al., 2015).

sample citations

Sample Citation: (Antipov and Pokryshevskaya 2012) (see: Antipov and Pokryshevskaya 2012, 33–35; also Antipov and Pokryshevskaya 2012, ch. 1 and *passim*)

A minus sign (-) before the @ will suppress mention of the author in the citation. This can be useful when the author is already mentioned in the text:

Antipov says blah (2012).

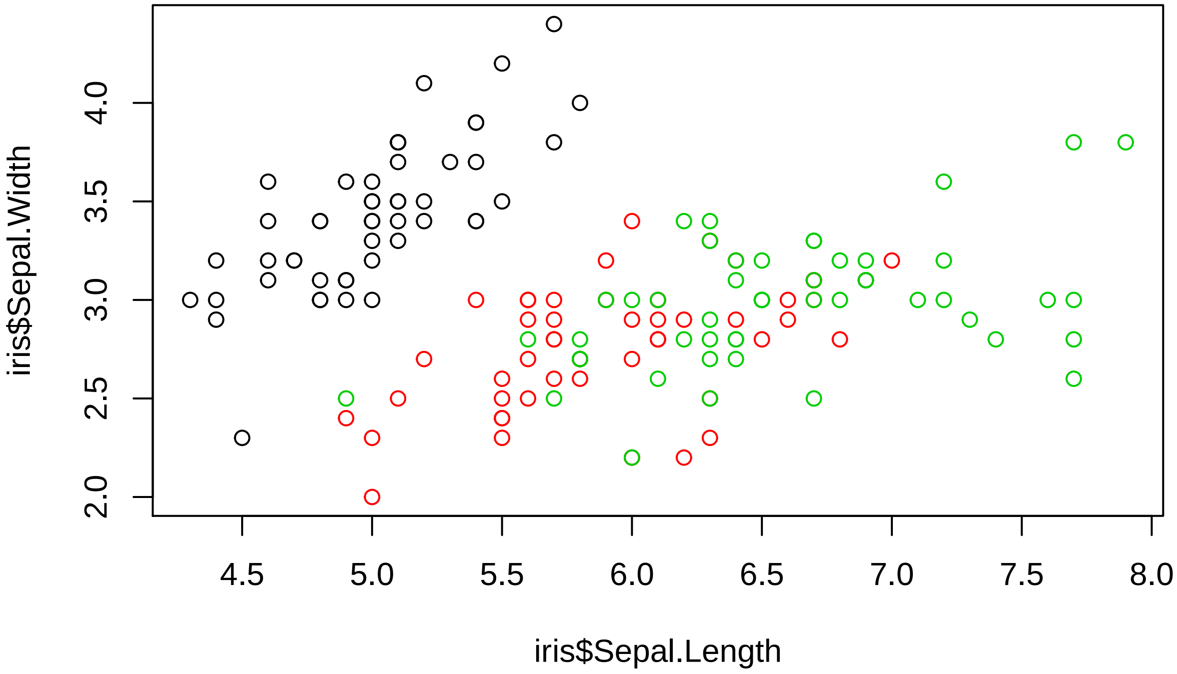
You can also write an in-text citation, as follows:

Antipov and Pokryshevskaya (2012) says blah.

Methodology

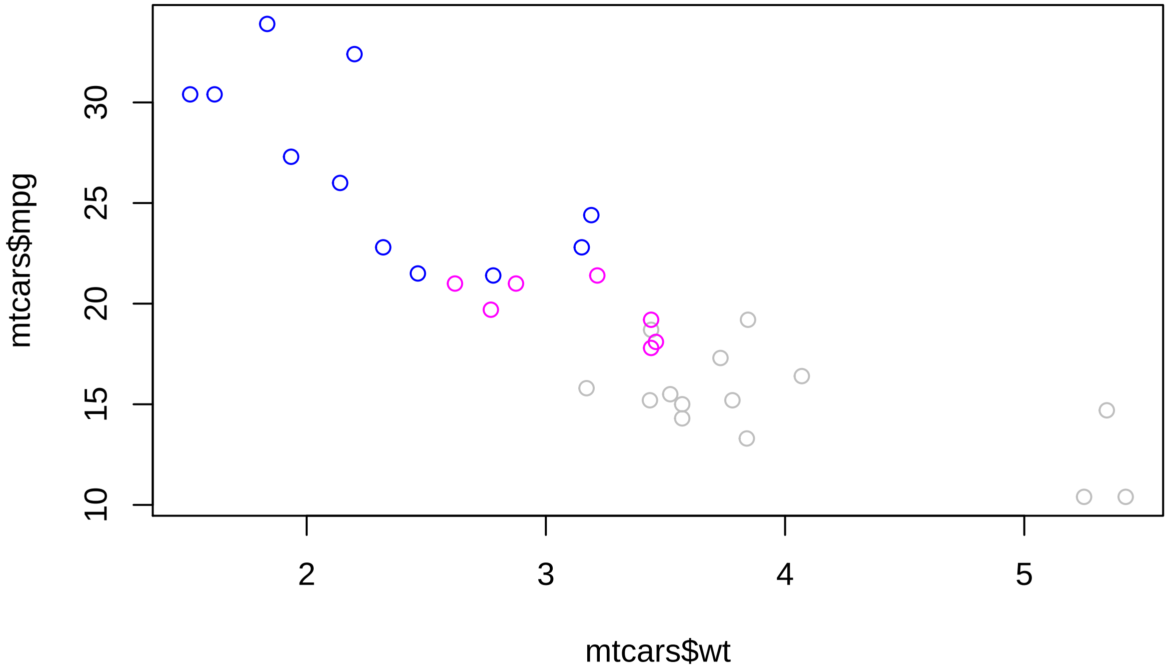
Methodology Section

```
plot(iris$Sepal.Length, iris$Sepal.Width, col = iris$Species)
```



Results

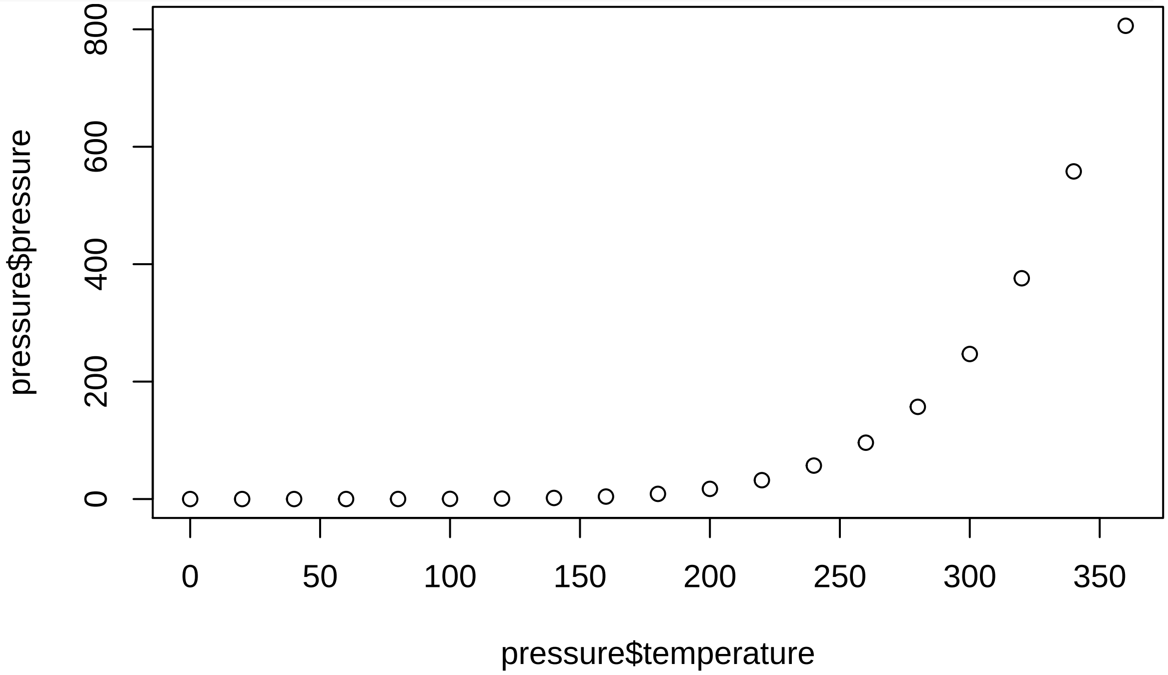
```
plot(mtcars$wt, mtcars$mpg, col = mtcars$cyl)
```



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Conclusions and Future Research

```
plot(pressure$temperature, pressure$pressure)
```



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

References

Antipov, Evgeny A., and Elena B. Pokryshevskaya. 2012. "Mass Appraisal of Residential Apartments: An Application of Random Forest for Valuation and a Cart-Based Approach for Model Diagnostics." *Expert Systems with Applications*.