Spatially Inspired Price Prediction for Car Rentals

ABSTRACT

The presence of spatial dependencies has often been ne- glected during the development of modern day machine learning models. In this project, i attempt to take into ac-count inherent *spatial* dependencies, with an ultimate goal to build a novel machine learning system to predict car rental prices using data from *Turo*, a peer to peer car-sharing com- pany.

KEYWORDS

turo, machine learning, XGBoost, random-forest, geo-spatial

1 INTRODUCTION

Prediction models have changed the way we estimate the value of everyday commodities. For instance, prediction mod- els using data from AirBnB, has led to many insights on apart- ment rental pricing, which, in the past was predominantly determined by human subjectivity. These prediction models often incorporate features which are directly related to ora part of the system on which predictions are being made.For instance, while predicting nighly rental prices in AirBnb apartments, features such as the number of bedrooms, thesize of the apartment, presence of a swimming pool, amongst others were considered.

Despite having seen widespread adoption and success,these classical prediction systems (e.g. AirBnb) ignored the presence of any spatial dependencies encoded implicitly or explicitly in the data. Encoding and using these spatial de-pendencies for predictive building prediction models, had led to the rise of spatially inspired approaches such as spatialregression, spatial clustering, and spatial random forest mod- els, which inherently encode spatial dependencies presentin data.

2 DATA COLLECTION AND PRE-PROCESSING

(A) Data Collection

The first dataset utilized was obtained from the kaggle profile of Christopher Lambert, and consisted of a 330MB JSONfile with a heavily nested format. The inital goal for data processing was to convert the data to a flat file format to facilitate further processing. Upon initiallyreading in the datasets using the Python Package Pandas, it was found that the dataset consisted of 36000 datapoints, each of which corresponded to a single car rental event. Given the size of the dataset, and the heavily nested nature of the data source,it was soon found that simple iteration and Pandas apply functions were impractical to flatten the dataset.

(B) Some Key Extracted and Engineered Features

Some of the key features extracted and engineeredfrom a combination of the two data sources were

(a) ID : The carID represents a unique ID for a par-ticular car in the dataset.

(b) Rating: The Rating represents the rating given by a particular user for the car for a particular trip.The rating field had some missing values, which were imputed using relevant strategies as outlined below: For every rental in the event that a rating was not provided, impute the missing value with the average rating for that renter's car's.

(c) Response Rate: The Response Rate is the number of times on average a host replied to a rental request (inpercent). In case no data was present, it was imputedwith a 0 percent.

(d) Listing Difference : This feature is an engineered feature, which represents the difference between thethe year in which a car was listed on the website, and the year in which the car was released. Intuitively,if this difference is high, it should ideally mean that either this car is a classic (high rates), or is a reallyold car at the end of its lifetime.

(e) Weekday, Month: Represents the day of the weekand the month on which the rental took place respec-tively. These 2 features are meant to detect trendsin weekly or seasonlity in car rental prices if any present.

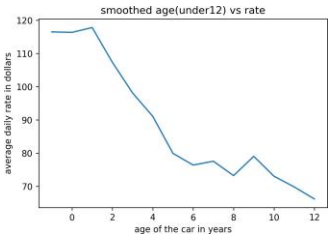
(f) Average Daily Price: The price for each particular car rental. This is our response variable

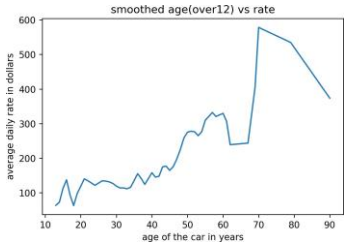
(C) Key Pre-processing Steps In addition to using hard-
ware acceleration, for unraveling (un-nesting the data), several pre-processing steps were undertaken. For in- stance, the time of listing of the car was provided inthe form of an EPOCH, and required the usage of the datetime package in Python in order to render and extract useful features from it. Similarly, all the lati-tude longitude coordinate pairs were converted into Geomtric (GIS) points using the *shapely* framework in Python using a traditional EPSG:4326 projection space.

3 EXPLORATORY ANALYSIS

(A) Common Attribute Exploration

The data set consisted of 36279 individual rentals onthe website Turo. Within the data set there were 58 different types of makes of cars, and 837 unique carmodels. The model years spanned from 1927 through 2019 for each car. The most expensive car in the data set was a Porsche 718 Cayman, costing $1999 to rentfor a day. On the other hand, the cheapest car to rent was a 2006 Kia Rio, or a 2011 Toyota Camry costing
$10 to rent for a day. Amongst the manufacturers, Toy-ota was the most common, having 4022 cars on Turo. The most uncommon car in the data set was a Yugo,having only 1 car listed. Amongst the cars listed, 94% of them were Automatic Transmission, and the rest had manual transmission. Within the listed vehicles,66% of them were sedans whereas 25% of them were SUV's, and 4% being trucks. The rest 3% were vans. On average sedans costed $96 a day, SUV's costed $108 a day, trucks costed $91 a day, and vans costed $72.5 a day.
Specifically onto one feature, we analyze the daily rateof the cars vs the age of them.





(A) Spacial Significance

Taking the dataset's Latitude and Longtitude attribute

and plotting them on a map, it is significant that Turo data points clusters major urban areas such as Los Angeles and New York City.

Geographical Distribution of Data Points

Random Forest Model :

----- Top Variable Importance ------

Variable	%
make	30
model	24
rentertriptaken	7
longitude	6
latitude	5
Listing difference	7

-------- Training Data ----------

R-Squared	0.87
RMSE	40.34

------- Validation Data ----------

R-Squared	0.65
RMSE	67.08

XGBoost Model :

----- Top Variable Importance ------

Variable	F-Score
model	940
month	624
Rating	456
weekday	452
latitude	447
longitude	382

-------- Training Data ----------

R-Squared	0.89
RMSE	44.64

------- Validation Data ---------

R-Squared	0.71
RMSE	57.12

The degree of correlation R^2 is not extremely high, but however, does suggest that the model is moderately good at predicting the price.
A good note, is that the month (an engineered feature), is the second most important feature selected by the XGBoost model. Additionally, model, make, month, and year of release of the car are the other determining factors in the Random Forest Regressor. This goes with our intuitive notion as well.