Yellow taxi trip regression

Nasser M Alqahtani
Mukhtar Al bin Hamad

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

Agenda

1- Introduction
2- Data
3- Algorithms
4- Tools
5- EDA
6- Models

Introduction

In project, build a machine learning regression model. The main purpose of this project is to provide predictions the price of trips.

Data

Taxi & Limousine Commission

Yellow taxi trip in NYC in July 2021.

+2.8M Observations

18 Features

Algorithms

Preparing the data.

Exploration the data and visualization.

Feature Engineering

converting categorical values to dummy.

Feature Selection

calculate the features correlation.

Methods

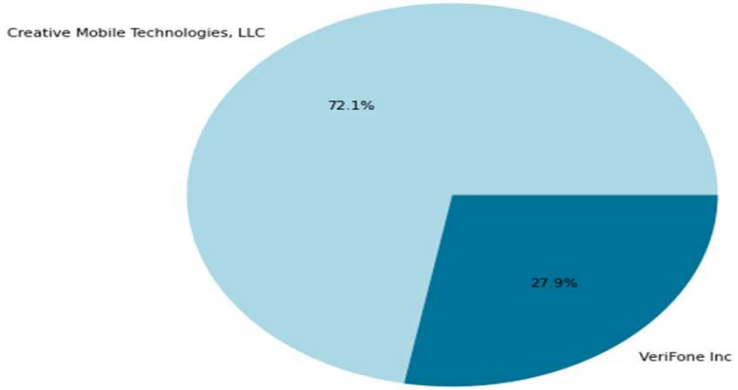
Linear regression, polynomial regression, ridge regression, lasso regression, ElasticNet, and Knn.

Tools

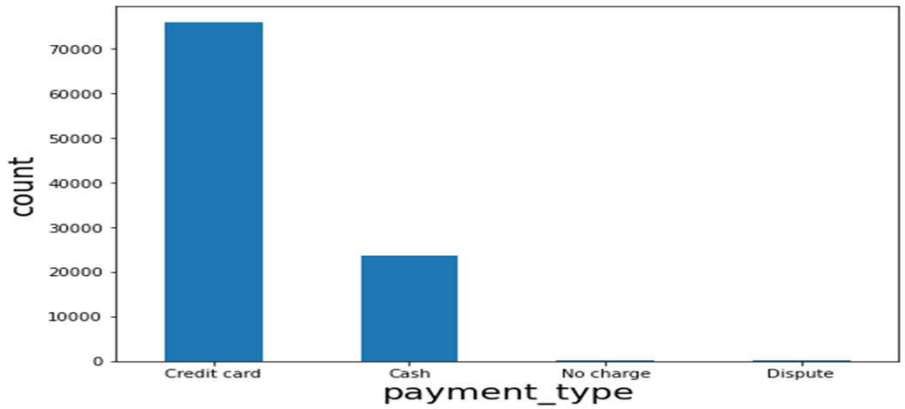
- **Python and Jupyter Notebook.**
- **Numpy and Pandas for data manipulation.**
- **Matplotlib and Seaborn for plotting visualization.**
- **Sklearn for ML algorithms.**

EDA

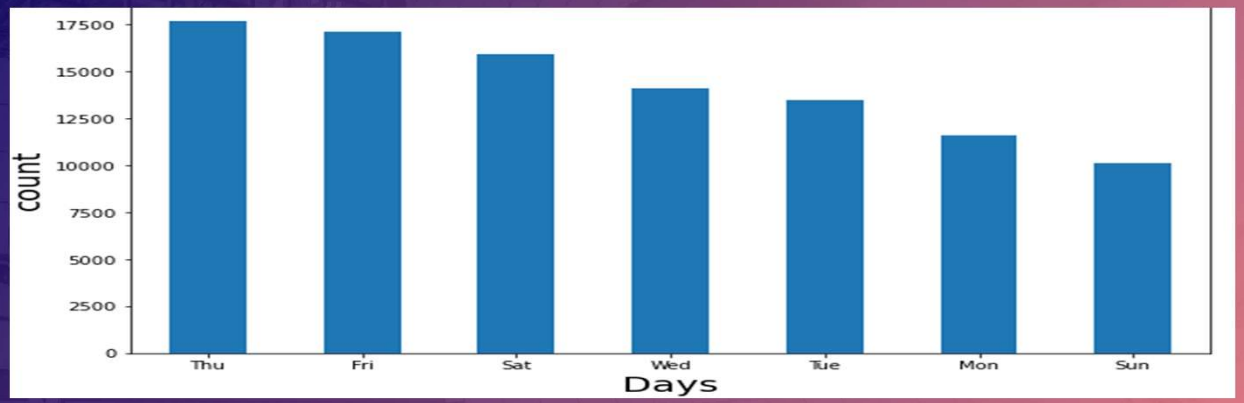
Distribution of VendorID



Show payment types



Distribution of trips in days



Model

Model \ R^2	Train	Validation	Test
Linear regression	0.771	0.746	0.779
Polynomial regression	0.872	0.890	0.845
Ridge regression	0.771	0.746	0.779
Lasso regression	0.771	0.746	0.776
Elastic Net regression	0.771	0.7460	0.776
KNN regression	0.914	0.919	0.865

Models

By applying the dataset on machine learning models as linear regression, polynomial regression, ridge regression, lasso regression, Elastic Net, and Knn, to predict the prices of the trips.

The best model	R^2 test	RMSE	MAE
KNN	0.865	5.45	2.48

R^2

ability of Model to fit the given data.

RMSE

Root-mean-square error.

MAE

Mean absolute error.

THANK YOU