دانشگاه صنعتی شریف

دانشکده مهندسی صنایع

پروژه درس برنامه ریزی حمل و نقل

نگارندگان:

آرین آقامحسنی، عمید نصیرپور، مرتضی وارسته

استاد درس:

جناب آقای دکتر عرفان حسن نایی

دستیاران آموزشی:

جناب آقای عرفان امانی بنی

پاییز ۱۴۰۳


بسم الله الرحمن الرحيم

فهرست

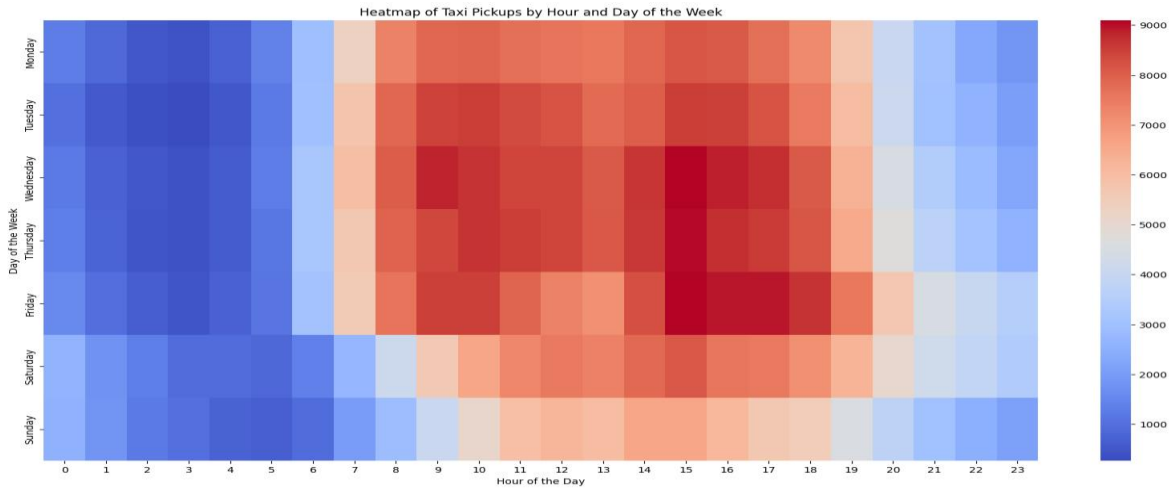
۱. خواسته اول (پیش پردازش داده ها):

در خواسته اول، از ما خواسته شده تعدادی پردازش را روی داده ها انجام دهیم. کد های مربوط به این بخش در فایل پروژه در بخش Task 1 با تفکیک هر بخش پیاده شده است.

۲. خواسته دوم (توصیف وو تفسیر داده ها):

۱.۲ نقشه گرمایی

شکل ۱ نشان دهنده نقشه گرمایی میزان استفاده از تاکسی برای هر روز و ساعت می‌باشد. همانطور که میبینیم میتوان روز را به طور کلی به سه قسمت متفاوت تقسیم کرد.مشاهده میشود که در تمام روز های هفته بطور تقریبی از ساعت ۱۲ شب تا ۶ صبح میزان استفاده از تاکسی ها پایین است، اما به مرور شدت یافته و از ساعت ۹ تا ۶ عصر بیشترین میزان را تجربه میکند.در این ساعات برای روز های مختلف هفته شدت استفاده از تاکسی های سبز در ساعات مختلف متفاوت است برای مثال در روز تعطیل یکشنبه ، میزان استفاده از تاکسی ها کاهش چشم گیری دارد ، در حالی که در روز های وسط هفته مانند چهارشنبه و پنجشنبه و جمعه ، میزان استفاده از تاکسی ها افزایش قابل توجهی میابد.

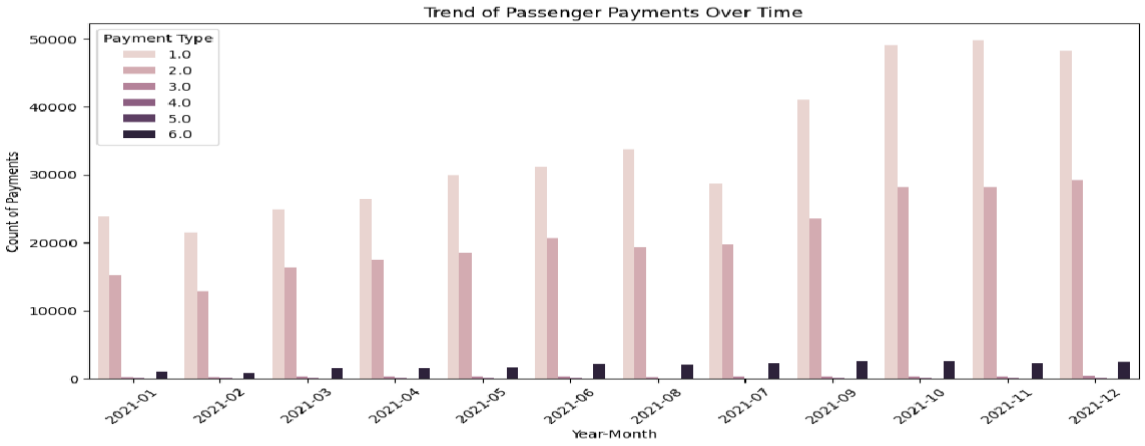


شکل ۱: نقشه گرمایی

۲.۲ نمودار های تحلیلی

۱.۲.۲ روند پرداخت مسافران در گذر زمان

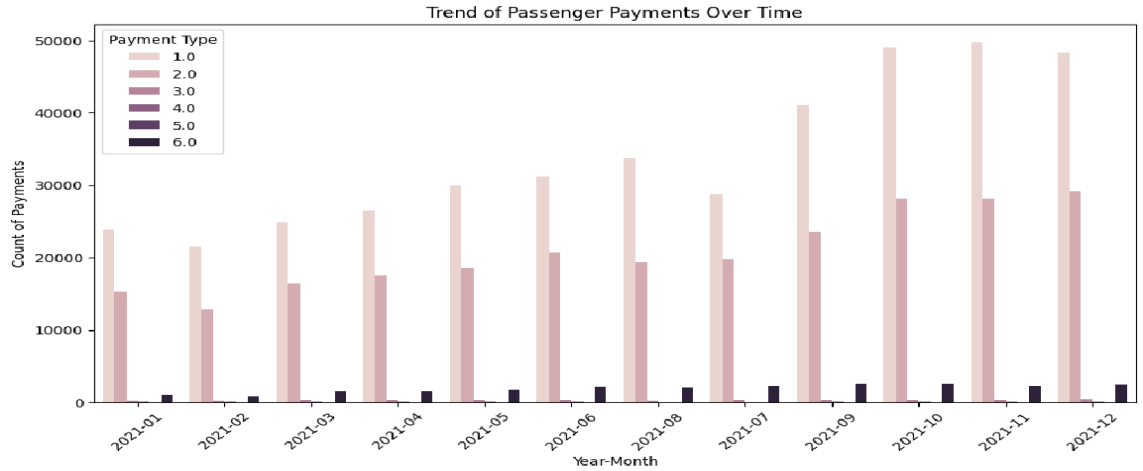
همانطور که در شکل۲ مشاهده میشود ، روند پرداخت با کارت اعتباری (نوع۱) در طول ۱۲ ماه گذشته رشد محسوسی داشته است. پرداخت نوع دوم نیز رفتار مشابهی داشته و با شیب کمتری در مجموع رشد کرده است. اما در کل نسبت به هم رفتار ثابتی داشتند. انواع باقی مانده از پرداخت نوسانات جزئی داشته و تغییرات محسوسی در آنها مشاهده نمیشود.



شکل۲: روند پرداخت مسافران در گذر زمان

۲.۲.۲ روند نوع سفر

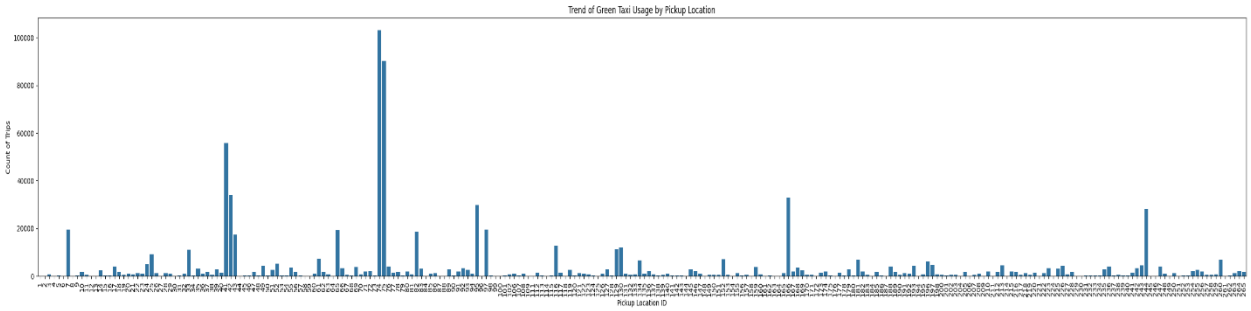
همانطور که در شکل۳ مشاهده میشود ، از ماه اول تا ماه دوازدهم سال ۲۰۲۱ ، نوع سفر۱ در مجموع رشد بیشتری داشته است و انواع باقی مانده سفر نوسانات جزئی داشته و رشد بسیار کمی را در میزان انجام این سفر ها شاهد هستیم.



شکل۳: روند نوع سفر

۳.۲.۲ روند استفاده از تاکسی های سبز در هر یک از بخش های شهر

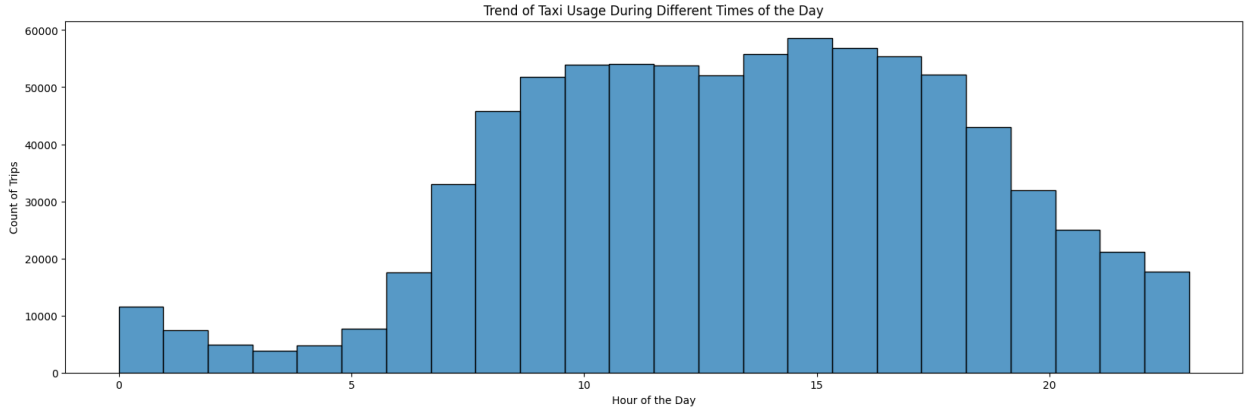
شکل۴ تعداد سفر هایی که در طول سال ۲۰۲۱ از لوکیشن آیدی های مربوطه آغاز شده است را نمایش میدهد.از طریق شکل میتوان متوجه شد که لوکیشن ها با شناسه های ۷۴و۷۵و۴۱و۱۶۶و۴۲و۲۴۴و۷ ، پر تردد ترین مکانها برای اغز سفر با تاکسی های سبز محل احتمالی ایستگاه های این نوع از تاکسی ها میباشند. (این نمودار در فایل کد مربوطه بهتر دیده میشود و قابل تحلیل است)



شکل ۴: روند استفاده از تاکسی های سبز در هر یک از بخش های شهر

۴.۲.۲ روند استفاده از تاکسی های سبز در بازه های زمانی روز

همانطور که در شکل۵ مشاهده میشود، در ساعات اولیه روز، سفر ها تعداد پایینی داشته و این میزان از ساعت ۶ صبح رو به افزایش می گذارد (با توجه به اینکه این ساعت مربوط به شروع روز کاری است نمودار منطقی هست) و روند افزایشی تا ساعت ۳ عصر ادامه میابد و پس از این ساعت مجددا شیب کاهشی به خود میگیرد (این کاهش نیز بعلت اتمام روز کاری است).



شکل ۵: روند استفاده از تاکسی های سبز در بازه های زمانی روز

۳. خواسته سوم (تحلیل دقیق داده ها):

از این بخش به بعد نیاز به داده های تمیز داریم، با توجه به اینکه در ۳ ستون مهم مقادیر خالی داریم، باید قبل از ادامه به آنها رسیدگی های لازم رو بکنیم. برای هر کدام از مسائل (با توجه به اینکه پر کردن مقادیر خالی درواقع یک تسک کلسیفیکیشن هست) باید یک مدل با داده هایی که داریم آموزش داده و آنرا روی داده هایی که مقادیر آنها را نداریم برای پیشبینی استفاده کنیم. (جدول ۱)

ستون مورد بررسی	درصد خالی بودن	مدل انتخابی	F1 score
Passenger Count	19.92%	RandomForestClassifier	0.87
Trip Tye	19.92%	RandomForestClassifier	0.99
Payment Type	16.34%	RandomForestClassifier	0.90

جدول ۱: پر کردن ستون های دارای مقادیر خالی

ستون trip_type بعلت اینکه تعداد خیلی زیادی از نوع ۱ دارد، مقداری آنبالاس بودن در مدل رو میتوان مشاهده کرد (و علت خیلی بالا بودن مدل نیز همین است)، برای رفع این مشکل در حالت کلی می توان از تکنیک های upsampling و یا downsampling و یا ترکیبی از این ۲ (SMOTE) استفاده کرد، اما با توجه به چولگی شدید داده ها این روش نیز کمکی نمیکند و صرفا مقدار کامپیوتیشن رو افزایش می‌دهد.

در مرحله بعد باید مشکل ستون هایی که کتگوریکال هستند رو با روش های انکودینگ حل کنیم، در این بخش برای هر ستون روشی مناسب با آن استفاده شده است که در جدول ۲ قابل مشاهده است.

ستون مورد بررسی	متد استفاده شده	توضیحات
Month	Label Encoding	با توجه به اینکه دیتا فقط برای ۱ سال است نیازی به انکودینگ سینوسی نداریم.
Day_of_week	One-Hot Encoding	منطقی نیست برای روز های هفته ترتیب عددی قائل شویم (جمعه با شنبه ۱ روز فاصله دارد نه ۷!)
Hour	Sine Cosine Transformation	این نوع انکودیگ برای ساعت بعلت این است که فاصله ها رو بطوری تبدیل میکند که مفهوم تکراری ساعت حفظ شود.
Trip_type	Label Encoding	-
Payment_type	One-Hot Encoding	علت مشابه روز هفته

جدول ۲: انکودینگ ستون های کتگوریکال

در نهایت میتوانیم به بخش ماتریس همبستگی برویم.

بازم به ذکر است که برای هر ستون توزیع ها چک شده و اگر مقادیری خارج از عرف بودند، از داده خارج شدند، اما با توجه به اینکه تعداد این موارد خیلی کم است، روش های outlier detection مانند zscore و IQR method نمیتوانند مفید باشند چرا که باعث حذف مقدار خوبی از داده (که درواقع خارج از عرف نیستند) می‌شوند.

۱.۳ ماتریس همبستگی

ماتریس همبستگی در شکل ۶ اطلاعاتی درباره روابط بین ویژگی‌ها و متغیر هدف total_amount ارائه می‌دهد. در اینجا برخی از مهم‌ترین مشاهدات آمده است:

۱. همبستگی‌های قوی با متغیر هدف (total_amount):

trip_distance: همبستگی مثبت قوی (۰/۸۳) با total_amount نشان می‌دهد که سفرهای طولانی‌تر معمولاً هزینه بالاتری دارند.

tip_amount: همبستگی مثبت متوسط (۰/۵۰) نشان می‌دهد که انعام‌های بیشتر با هزینه کل بالاتری همراه هستند.

tolls_amount: همبستگی مثبت ضعیف (۰/۲۳) نشان‌دهنده تأثیر جزئی عوارض جاده‌ای بر هزینه کل است.

passenger_count: همبستگی مثبت متوسط (۰/۷۵) نشان می‌دهد که سفرهایی با تعداد مسافر بیشتر معمولاً هزینه کل بالاتری دارند.

trip_type: همبستگی مثبت قوی (۰/۸۳) با total_amount نشان می‌دهد که انواع مختلف سفر می‌توانند دسته‌بندی‌های قیمتی متفاوتی داشته باشند.

۲. ویژگی‌های با همبستگی منفی با total_amount:

VendorID: همبستگی منفی ضعیف (-۰/۱۵) نشان می‌دهد که نوع فروشنده ممکن است تأثیری جزئی بر هزینه کل داشته باشد.

month: همبستگی منفی ضعیف (-۰/۱۱) نشان‌دهنده کاهش جزئی هزینه‌ها در ماه‌های خاص (احتمالاً زمستان).

۳. روزهای هفته:

همبستگی‌های روزهای هفته: روزهایی مانند day_of_week_Monday و day_of_week_Sunday همبستگی منفی متوسط با هزینه کل دارند، که نشان‌دهنده تفاوت در تقاضا در روزهای هفته است.

۴. ویژگی‌های زمانی:

hour_sin و hour_cos: این ویژگی‌ها همبستگی زیادی با trip_duration (۰/۸۲ و ۰/۳۱) دارند و نشان می‌دهند که زمان روز بر مدت زمان سفر و به تبع آن هزینه تأثیر دارد.

hour: همبستگی ضعیف (۰/۱۷) با total_amount دارد، اما تبدیل‌های دایره‌ای آن (hour_sin و hour_cos) بهتر مدل‌سازی می‌کنند.

۵. نوع پرداخت:

نوع پرداخت: ویژگی‌های مربوط به نوع پرداخت همبستگی‌های بالایی دارند (از ۰/۸۹ تا ۱/۰)، که نشان‌دهنده تأثیر قوی نوع پرداخت بر هزینه کل است.

۶. ملاحظات چندهمبستگی:

ویژگی‌هایی مانند trip_type، passenger_count، trip_distance و tip_amount همبستگی‌های بالایی با هدف دارند و ممکن است در پیش‌بینی هزینه کل تأثیر زیادی داشته باشند.

ویژگی‌های دایره‌ای مانند hour_sin و hour_cos همبستگی‌های قوی با hour دارند، بنابراین نیاز به انتخاب ویژگی یا منظم‌سازی برای جلوگیری از بیش‌برازش داریم.

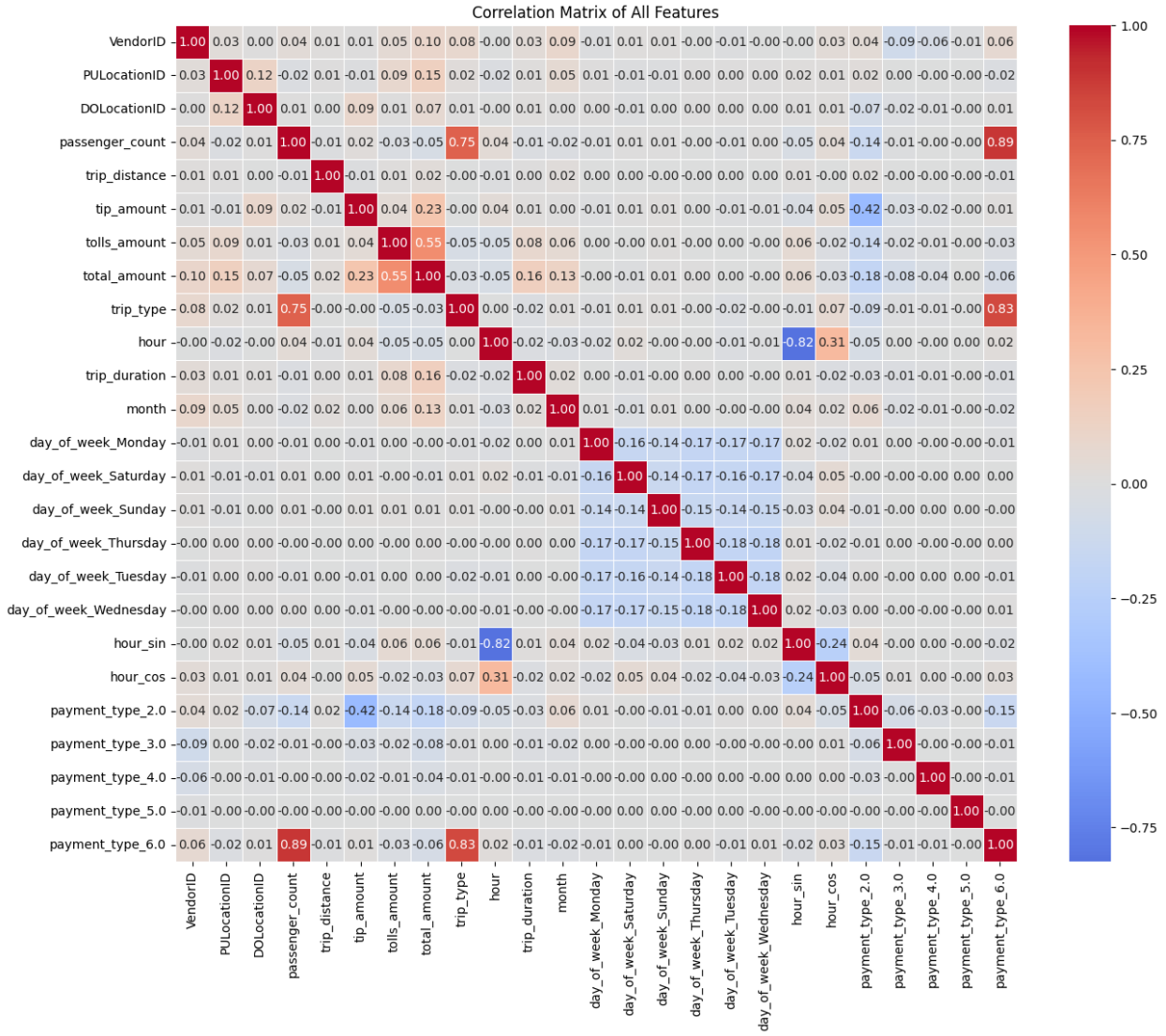
در نتیجه داریم :

trip_distance و trip_type (همبستگی مثبت قوی)

tip_amount و tolls_amount (همبستگی مثبت متوسط)

passenger_count (همبستگی مثبت متوسط)

کاهش چندهمبستگی با ترکیب ویژگی‌های همبسته یا استفاده از روش‌های منظم‌سازی می‌تواند مفید باشد. همچنین، ویژگی‌های زمانی و نوع پرداخت نیز نقش مهمی دارند و باید در مهندسی ویژگی‌ها در نظر گرفته شوند.



شکل ۶: ماتریس همبستگی داده ها

۲.۳ انتخاب متغیر های تصمیم گیری

با توجه به اینکه مقدار داده ها نسبتا زیاد است و تعداد ویژگی ها زیاد، روش backward نمیتواند کمک زیادی بکند (اما پیاده سازی شده است)

آزمون مبع کای برای داده های کتگوریکال طراحی شده اما اینجا با توجه به اینکه بعضی از ستون ها مقادیر پیوسته دارند بهترین مدل برای استفاده نیست.

روش های forward و Random Forest بهترین ها هستند.

نتایج همه این روش ها روی داده ها در جدول ۳ نشان داده شده.

Feature	Chi-Square Test	Random Forest	Backward Feature Selection (RFE)	Forward Feature Selection
VendorID	Yes	Yes	Yes	Yes
PULocationID	Yes	Yes	No	No
DOLocationID	Yes	Yes	No	No
passenger_count	Yes	Yes	Yes	No
trip_distance	No	Yes	No	Yes
tip_amount	Yes	Yes	Yes	Yes
tolls_amount	Yes	Yes	Yes	Yes
trip_type	Yes	Yes	Yes	Yes
hour	Yes	Yes	No	Yes
month	Yes	Yes	Yes	Yes
hour_sin	Yes	Yes	Yes	No
hour_cos	Yes	Yes	Yes	Yes
payment_type_2.0	Yes	Yes	Yes	Yes
payment_type_3.0	Yes	No	Yes	Yes
payment_type_4.0	Yes	No	Yes	Yes
payment_type_5.0	No	No	Yes	No
payment_type_6.0	Yes	Yes	Yes	Yes
day_of_week_Monday	No	Yes	Yes	No
day_of_week_Saturday	No	No	Yes	No
day_of_week_Sunday	No	No	Yes	No
day_of_week_Tuesday	No	No	Yes	No
day_of_week_Thursday	No	No	Yes	No
day_of_week_Wednesday	No	No	Yes	No

جدول ۳: ویژگی های منتخب

با توجه به این نتایج برای هر یک از تسک های رگرشن و کلسیفیکشن ویژگی های متفاوتی استفاده شده که در بخش های هر کدام به آنها اشاره میکنیم.

متد دیگری که در این بخش استفاده شده، متد VIF (Variance Inflation Factor) است، این متد با مقدار بهرانی ۱۰ نتوانست هیچ ویژگی ای را حذف کند.

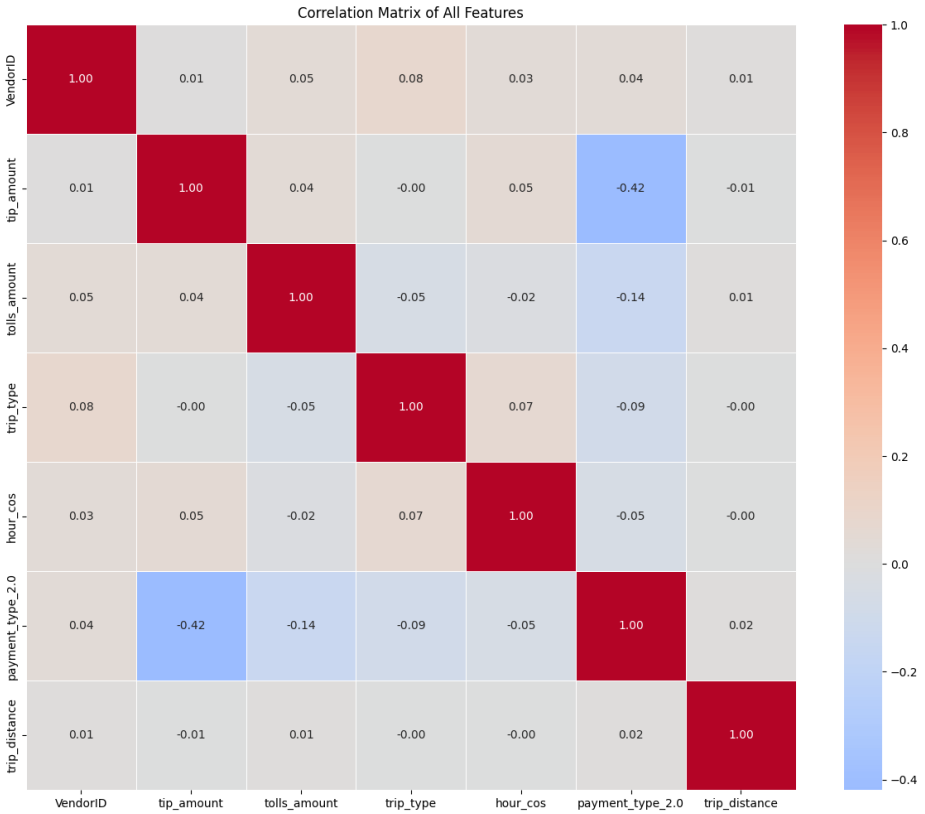
روش دیگری که در این بخش استفاده شد، VAE (Variational Auto Encoder) ها هستند. این مدل های عمیق با داده هایی که به آنها داده میشود آن هارو به بعدی پایین تر میبرد (Latent Space) و در واقع نمایشی جدید از داده ها می باشد. این روش بطور محسوس ویژگی های خاصی را انتخاب نمیکند بلکه بهترین ترکیب از همه ویژگی ها را در تعداد محدودی ارائه می‌دهد. توضیحات بیشتر این روش در نوت بوک پروژه درس هست.

۳.۳ مدل پیشبینی پرداخت یاعدم پرداخت انعام

در مرحله اول در این بخش باید ویژگی های استفاده برای مدل را بیان کنیم، این ویژگی ها عبارتند از :

```
"VendorID", "tip_amount", "tolls_amount", "trip_type", "hour_cos", "payment_type_2.0", "trip_distance"
```

و مورد آخری که باید چک شود ماتریس همبستگی است :



همانطور که میبینیم هیچ کدام از ویژگی ها وابستگی بالایی به متغیر هدف ما ندارد. پس میتوانیم ادامه دهیم.

مدل هایی که برای این بخش انتخاب کردیم عبارتند از :

RandomForestClassifier, Logistic Regression, XGBoost Classifier, AdaBoosts Classifier, Neural Networks, CatBoost and High Grad

و برای هر یک، گرید سرچ برای هایپرپارامتر هارا نیز در بخش مربوط به همان مدل پیاده سازی کردیم.

تنها توضیحی که باقی میماند، موضوعاتی است که باید به آنها توجه میکردیم:

۱. دیتای ولیدیشن و تست قبل شروع مدل سازی جدا شدند و در نتیجه دیتا لیکیج نخواهیم داشت.
۲. برای دسته بندی هدف (انعام دادن یا ندادن) متد SMOTE پیاده سازی شده تا مدل به یکی از ۲ حالت چولگی نداشته باشد.
۳. برای جلوگیری از اور فیت شدن مدل، ارتباط ولیدیشن اسکور و ترین اسکور بطور مخصوص مورد بررسی قرار گرفته

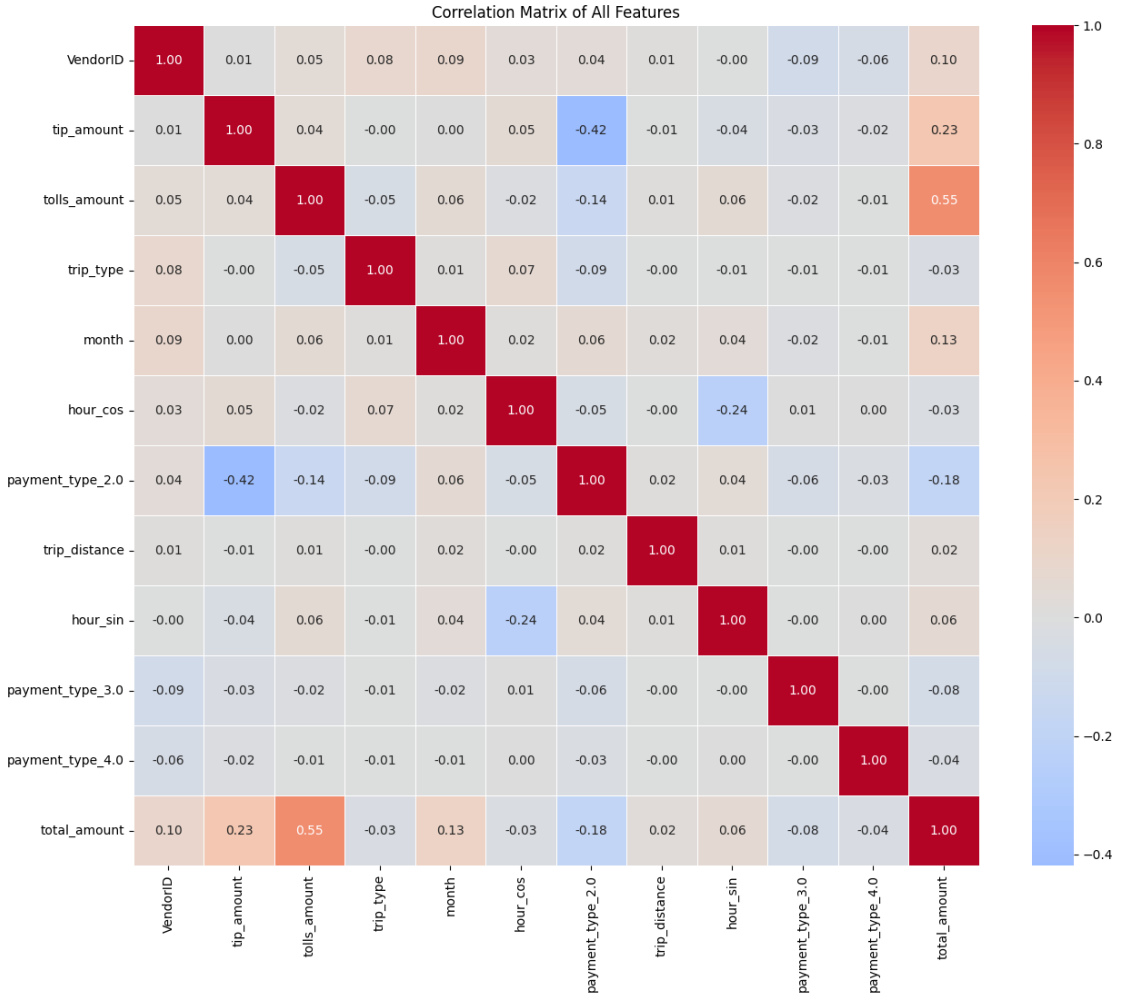
در نهایت و بعد از پیاده سازی مدل ها، مدلHigh Grad (با f1 score = 90 و ROC = 0.9497) بهترین نتیجه را داشت (البته مدل های شبکه عصبی و XGBoost بطور شدیدا نزدیکی به آن دوم و سوم بودند)

۳.۳ مدل پیشبینی مقدار کل پرداختی

در مرحله اول در این بخش باید ویژگی های استفاده برای مدل را بیان کنیم، این ویژگی ها عبارتند از :

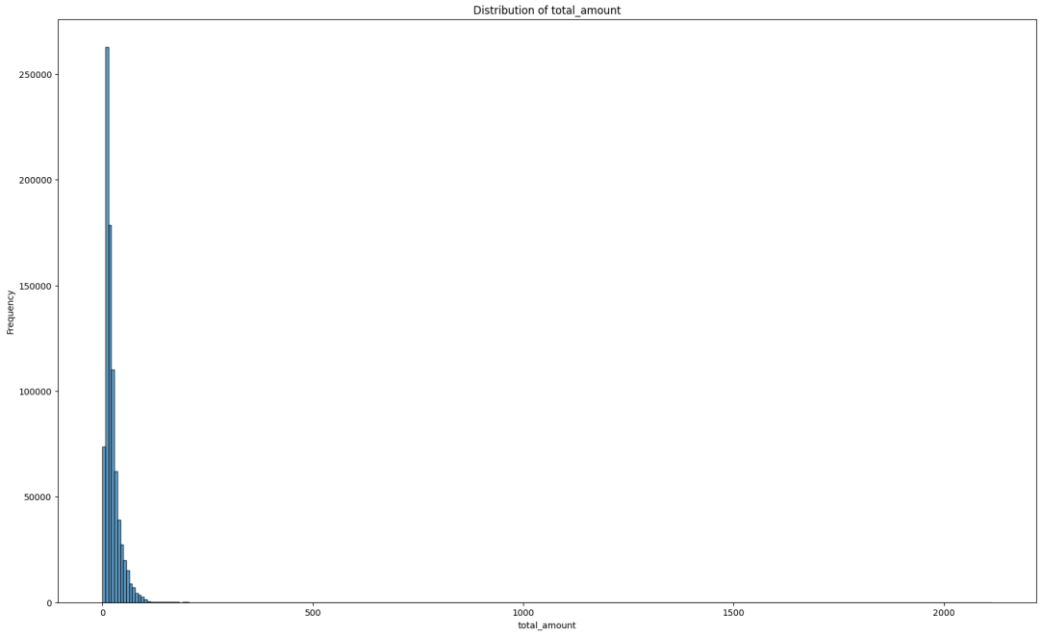
```
'VendorID', 'tip_amount', 'tolls_amount', 'trip_type', 'month',
'hour_cos', 'payment_type_2.0', 'trip_distance', 'hour_sin',
'payment_type_3.0', 'payment_type_4.0', 'total_amount'
```

و مورد بعدی که باید چک شود ماتریس همبستگی است :



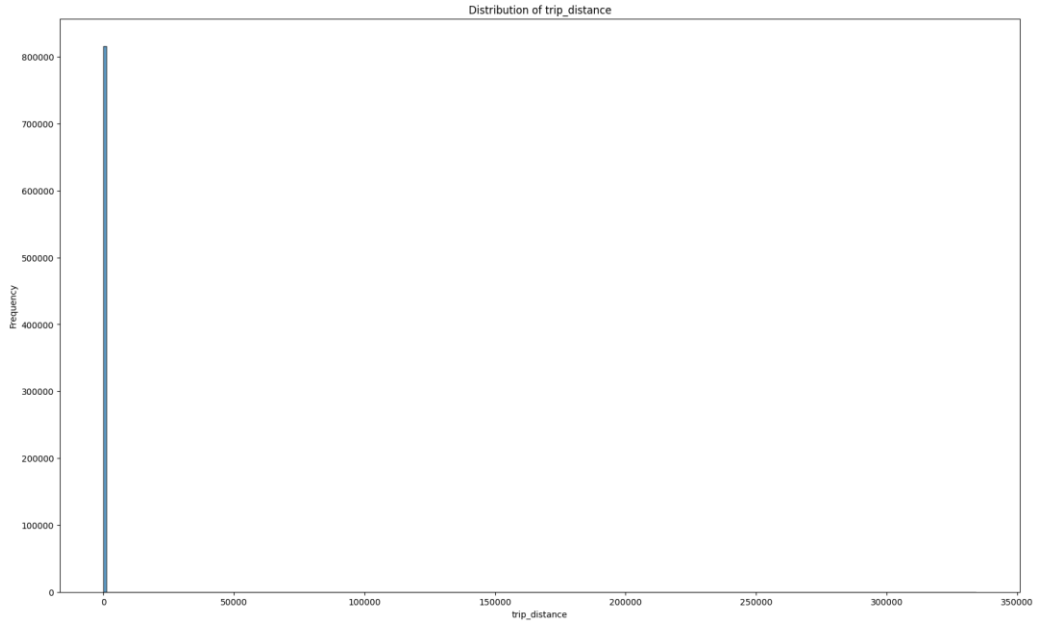
همانطور که میبینیم هیچ کدام از ویژگی ها وابستگی بالایی به متغیر هدف ما ندارد. پس میتوانیم ادامه دهیم.

مورد آخری که برای پیش پردازش مدل ها پیاده کردیم، استفاده از transformation ها است. در این پروژه توزیع هر کدام از ستون ها چک شده و در صورت نیاز تبدیلی مطابق با آن پیاده سازی شده. توزیع ستون هدف (total_amount) در نمودار زیر آماده :

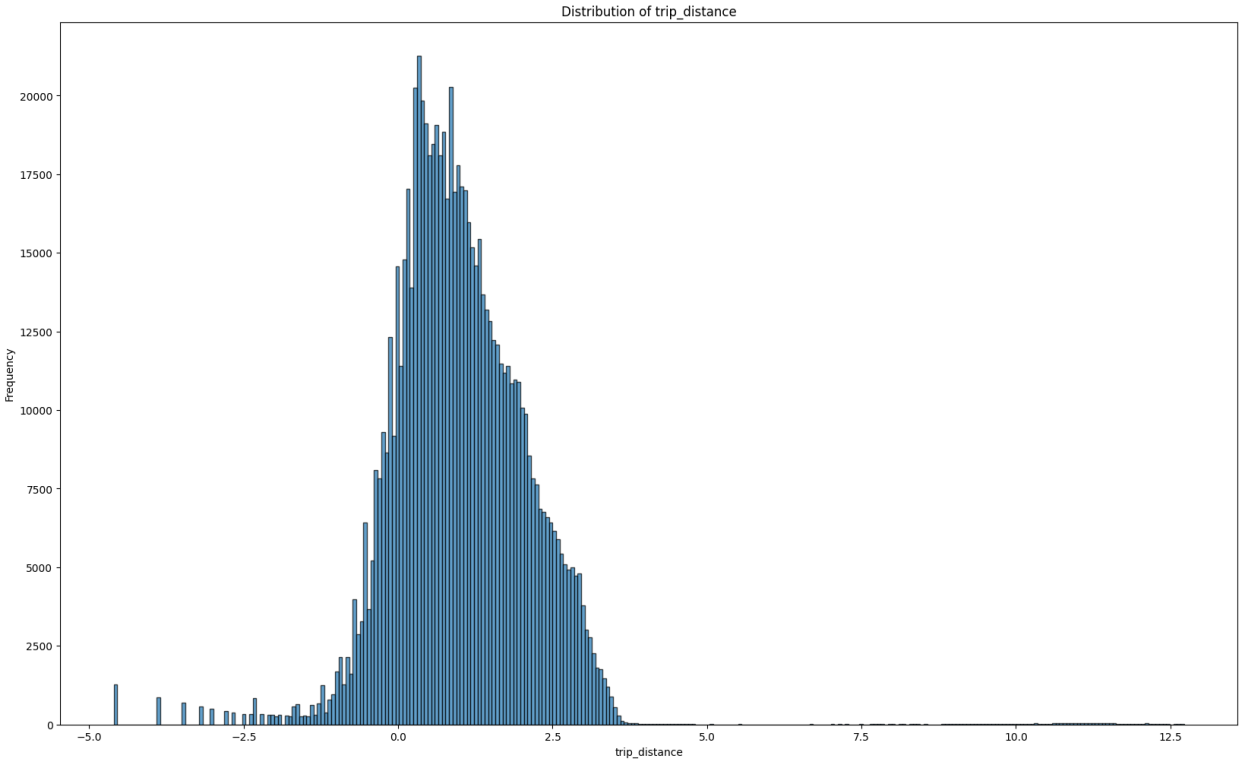


این توزیع مناسب بوده و نیازی به ترنسفورمیشن های ما ندارد.

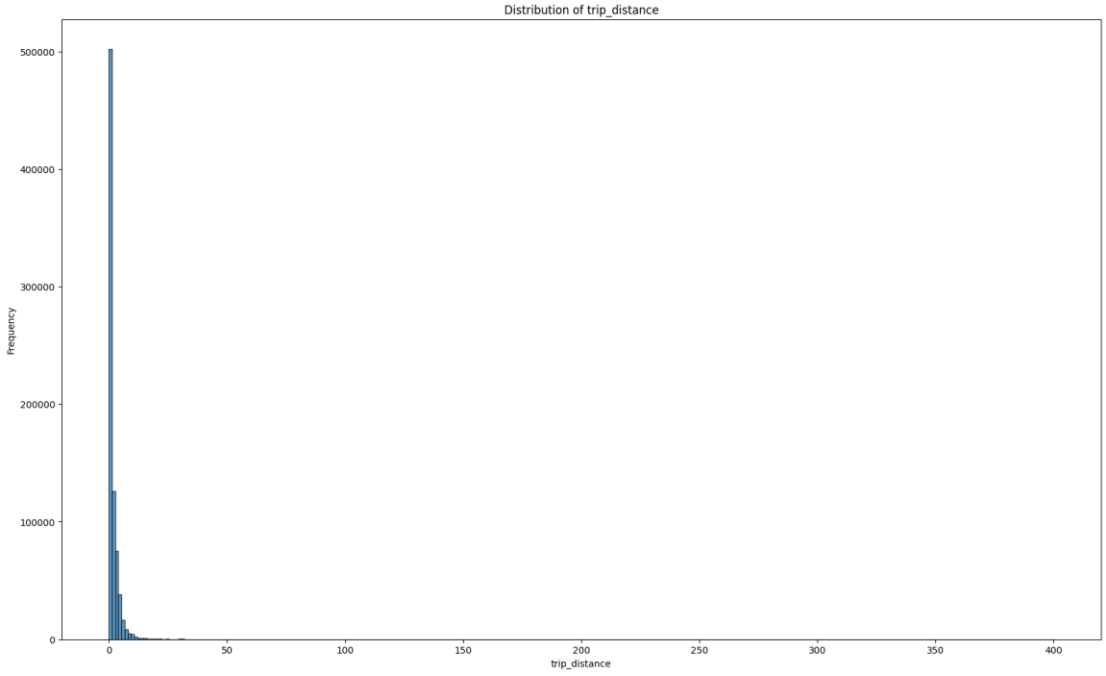
اما ویژگی trip_distance (که از ویژگی هایی است که ما مهندسی کردیم) میتواند کاندید مناسبی برای ترنسفور شدن باشد:



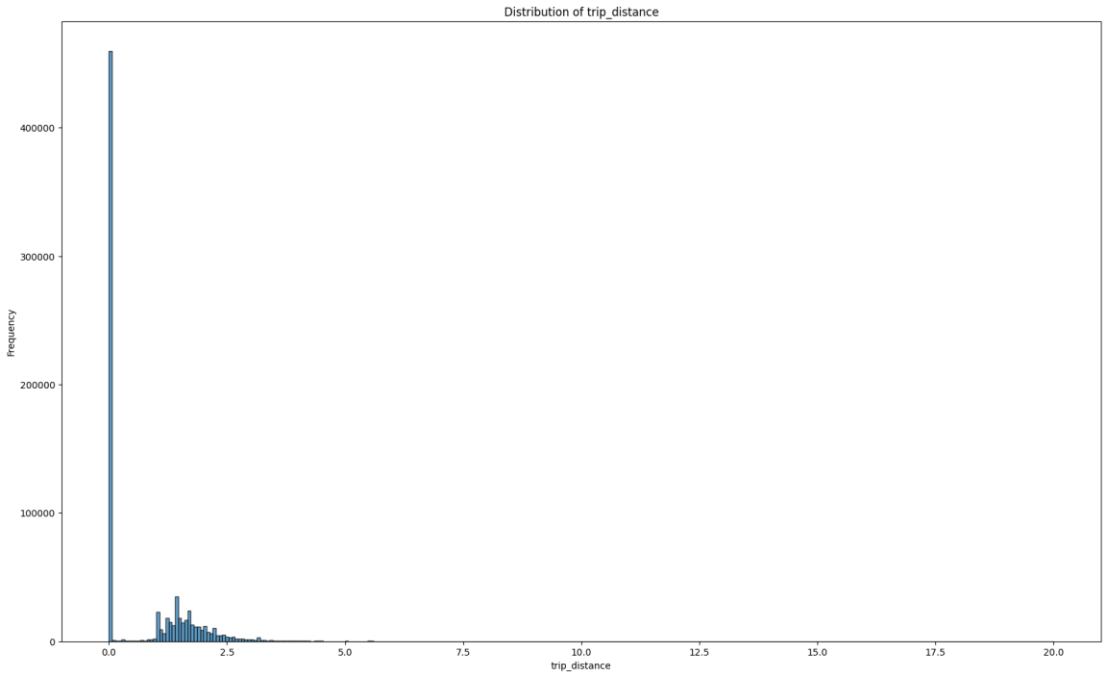
این توزیع برای مدل های رگرسیون مناسب نمیباشد، پس با استفاده از تبدیل لگاریتم آن را به نمودار زیر تبدیل میکنیم :



برای ستن tip_amount هم داریم :



بعد از تبدیل رادیکل :



که تغییر چندای نیست اما میتواند به مدل ها کمک کند.

بعد از این به بخش مدل سازی میرسیم

مدل هایی که برای این بخش انتخاب کردیم عبارتند از :

RandomForestRegressor, XGBoost Regressor, AdaBoost, Neural Networks, CatBoost and High Grad

و برای هر یک، گرید سرچ برای هایپرپارامتر هارا نیز در بخش مربوط به همان مدل پیاده سازی کردیم.

تنها توضیحی که باقی میماند، موضوعاتی است که باید به آنها توجه میکردیم:

۱. دیتای ولیدیشن و تست قبل شروع مدل سازی جدا شدند و در نتیجه دیتا لیکیج نخواهیم داشت.
۲. برای جلوگیری از اور فیت شدن مدل، ارتباط ولیدیشن اسکور و ترین اسکور بطور مخصوص مورد بررسی قرار گرفته

در نهایت و بعد از پیاده سازی مدل ها، مدل RandomForestRegressor (با R2 Score = 0.8192) بهترین نتیجه را داشت (البته مدل های CatBoost و XGBoost دوم و سوم بودند)