گزارش تکلیف پنجم درس الگوریتم های علوم داده

نام و نام خانوادگی: فاطمه ترودی

شماره دانشجویی: ۴۰۳۴۲۲۰۴۸

نام استاد: دکتر سعیدرضا خردپیشه

نیمسال دوم ۰۴-۱۴۰۳

فهرست مطالب

۱ .مقدمه و خلاصه‌ای از پروژه

این پروژه با هدف ساخت مدل‌هایی برای تخمین هزینه بیمه‌نامه، با استفاده از داده‌های مشتریان و اطلاعات بیمه‌نامه، طراحی شده است. مسیر کاری که دنبال شده، یک فرآیند استاندارد در یادگیری ماشین است:

۱. بارگذاری و بررسی داده‌ها
۲. رسیدگی به داده‌های گمشده و مهندسی ویژگی
۳. کدگذاری متغیرهای دسته‌بندی
٤. آموزش و ارزیابی مدل‌ها(Random Forest) ، XGBoost، Linear Regressionو مدل‌های با regularized مانند Ridge و(Lasso
٥. پیش‌بینی روی داده‌های آزمایشی و ذخیره نتایج

از آنجایی که پیش‌بینی هزینه بیمه‌نامه یک **مسئله رگرسیون** (خروجی پیوسته) است، از چندین الگوریتم رگرسیون استفاده شده. برای ارزیابی عملکرد مدل‌ها از معیارهایی مانند **میانگین مربعات خطا (MSE)** استفاده شده و برای بهبود دقت، **متغیر هدف (هزینه بیمه‌نامه)** با لگاریتم تبدیل شده است.

---

۲ .بارگذاری و بررسی اولیه داده‌ها

در ابتدا، داده‌های آموزش و آزمون از فایل‌های CSV بارگذاری شدند. بررسی‌های اولیه شامل مشاهده ابعاد، نگاهی به چند سطر اول و شناخت نوع داده‌ها در هر ستون است. یک گام مهم در این مرحله، حذف ستون **ID**است، زیرا این ستون تنها یک شناسه بوده و هیچ اطلاعات پیش‌بینی‌کننده‌ای در خود ندارد.

علاوه بر این، ستون **"Coverage Commencement"**که زمان شروع پوشش بیمه را نشان می‌دهد، به فرمت تاریخ و زمان (datetime)تبدیل شد تا بتوان ویژگی‌های زمانی مانند سال و ماه را از آن استخراج کرد.

---

۳ .رسیدگی به داده‌های گمشده و مهندسی ویژگی

وجود مقادیر گمشده (Missing Values) می‌تواند باعث بروز مشکل در برخی مدل‌ها، به‌ویژه رگرسیون خطی، شود. تابعی برای مدیریت این مقادیر تعریف شد که رویکرد آن بر اساس نوع داده‌ها متفاوت است:

- برای داده‌های شبیه به دسته‌بندی، از **پر کردن با مد (Mode Imputation)** استفاده شد.
- برای ویژگی‌های پیوسته، از **میانگین (Mean Imputation)** استفاده شد.
- برای برخی ویژگی‌های دسته‌بندی گمشده، یک دسته جدید به نام **"Unknown"**اضافه شد.
- مقادیر گمشده در **"Yearly Earnings"**(درآمد سالانه) با **میانه (Median)** پر شدند تا تأثیر درآمدهای بسیار بالا یا پایین را خنثی کنند.
- برای **"Dependent Count"**(تعداد افراد تحت تکفل)، مقادیر گمشده به **۱** -تبدیل شدند.

مهندسی ویژگی نیز بخش مهمی از این مرحله بود. ویژگی‌های جدیدی مانند نسبت‌ها (نسبت درآمد به تعداد افراد تحت تکفل)، **ویژگی‌های تعاملی** (مانند حاصل‌ضرب تعداد ادعاها در خودرو) و تبدیلات لگاریتمی (Logarithmic) ساخته شدند. این ویژگی‌ها بر اساس اهمیت آن‌ها که توسط مدل‌های **Random Forest** و **XGBoost** مشخص شده، در مدل‌ها گنجانده شدند تا قدرت پیش‌بینی را افزایش دهند.

---

۴ .کدگذاری متغیرهای دسته‌بندی

اغلب مدل‌های یادگیری ماشین نیاز به ورودی عددی دارند. به همین دلیل، متغیرهای دسته‌بندی (مانند "شغل"، "وضعیت تأهل") با استفاده از روش **One-hot Encoding** به اعداد تبدیل شدند. این روش هر ستون دسته‌بندی را به چندین ستون باینری (۰ و ۱) تبدیل می‌کند تا مدل‌ها بتوانند آن‌ها را به طور موثر پردازش کنند.

---

۵ .مدل Random Forest

اولین مدلی که آموزش داده شد، یک **Random Forest Regressor** است. این مدل، یک مدل **ensemble** بوده که از ترکیب چندین درخت تصمیم‌گیری ساخته شده تا دقت و مقاومت در برابر نویز را افزایش دهد .Random Forest قادر به مدل‌سازی روابط غیرخطی بوده و با داده‌های با ابعاد بالا به خوبی کار می‌کند.

در این مرحله، متغیر هدف لگاریتمی شد و سپس پیش‌بینی‌ها به حالت اولیه بازگردانده شدند .**اهمیت ویژگی (Feature Importance)** نیز با استفاده از Permutation Importance ارزیابی شد تا مشخص شود کدام ویژگی‌ها بیشترین تأثیر را در پیش‌بینی‌ها دارند.

---

۶ .مدل XGBoost

رویکرد بعدی، استفاده از **XGBoost (Extreme Gradient Boosting)** بود. این الگوریتم به دلیل دقت و کارایی بالا شناخته شده است XGBoost .درخت‌ها را به صورت متوالی می‌سازد، به طوری که هر درخت جدید خطاهای درخت‌های قبلی را تصحیح می‌کند. پارامترهای مختلفی مانند `gamma` و `learning_rate` به تنظیم و جلوگیری از **بیش‌برازش** کمک می‌کنند.

اهمیت ویژگی در این مدل نیز با استفاده از توابع داخلی XGBoost به صورت بصری نمایش داده شد.

---

۷ .رگرسیون خطی و مدل‌های با Regularization

علاوه بر مدل‌های پیچیده، از مدل‌های رگرسیون خطی نیز استفاده شد تا به عنوان یک **خط پایه (Baseline)** عمل کنند و نتایج آن‌ها با مدل‌های پیچیده‌تر مقایسه شود .مدل‌های مورد استفاده عبارتند از:

- **رگرسیون خطی :(Ordinary Linear Regression)** یک مدل ساده که رابطه خطی را فرض می‌کند.
- **:Ridge Regression (L2)**با کوچک کردن ضرایب، به کاهش بیش‌برازش کمک می‌کند.
- **:Lasso Regression (L1)**علاوه بر کاهش ضرایم، برخی از آن‌ها را به صفر می‌رساند و در واقع یک نوع انتخاب ویژگی (Feature Selection) را انجام می‌دهد.
- **:ElasticNet**ترکیبی از Ridge و Lasso است تا تعادلی بین کاهش ضرایب و انتخاب ویژگی برقرار کند.

---

۸ .پیش‌بینی و خروجی نهایی

پس از آموزش مدل‌ها، از آن‌ها برای پیش‌بینی روی داده‌های آزمایشی استفاده شد. تمامی مراحل پیش‌پردازش (preprocessing)که بر روی داده‌های آموزش اعمال شدند، به طور مشابه بر روی داده‌های آزمون نیز پیاده‌سازی شدند. در نهایت، نتایج پیش‌بینی‌ها در قالب فایل‌های CSV ذخیره شدند که این کار، قابلیت تکرارپذیری فرآیند و مقایسه آسان مدل‌ها را فراهم می‌کند.