

১. ভূমিকা ও প্রজেক্টের প্রেক্ষাপট

যদিও একজন জুনিয়র ডেটা সায়েন্টিস্টকে সাধারণত ডেটা প্রসেসিং বা মডেল ট্রেনিংয়ের মতো ছোট কাজ দেওয়া হয়, কিন্তু যখন কেউ লিডারশিপ পজিশনে যান, তখন তাঁকে শুরু থেকে একটি সমস্যা সমাধানের জন্য সঠিক ধাপগুলি কী হবে, তা নির্ধারণ করতে হয়। একজন ডেটা সায়েন্টিস্টের অগ্রগতি নির্ভর করে তাঁর সমস্যা সমাধানের দক্ষতার ওপর।

- বিজনেস কেস স্টাডি: একটি বাস্তব উদাহরণ হিসেবে নেটফ্লিক্স (**Netflix**)-এর উদাহরণ নেওয়া হয়েছে।
- বিজনেস প্রবলেম: নেটফ্লিক্সের মিটিংয়ের মূল আলোচ্য বিষয় হলো রাজস্ব (**Revenue**) কীভাবে বাড়ানো যায়। রাজস্ব বাড়ানোর তিনটি সম্ভাব্য উপায়: নতুন গ্রাহক আনা, পুরোনো গ্রাহকের কাছ থেকে বেশি চার্জ করা, অথবা যারা প্ল্যাটফর্ম ছেড়ে চলে যাচ্ছেন (**Churn**) তাদের থামানো।
- সিদ্ধান্ত: দল সিদ্ধান্ত নেয় যে, রাজস্ব বাড়ানোর সবচেয়ে লাভজনক পথ হলো গ্রাহকদের প্ল্যাটফর্ম ছেড়ে যাওয়া (**Churn Rate**) কমানো।

২. প্রথম ধাপ: ব্যবসায়িক সমস্যাকে গাণিতিক সমস্যায় রূপান্তর

একটি সফল ML প্রজেক্ট শুরু হয় ব্যবসায়িক লক্ষ্যকে সংখ্যায় (metrics) প্রকাশ করার মাধ্যমে।

- গাণিতিক লক্ষ্য: প্রজেক্টের লক্ষ্য স্থির করা হয়: বর্তমানে ৪% থাকা চর্ন রেটকে ৬ মাসের মধ্যে কমিয়ে ৩.৭৫%-এ নিয়ে আসা।

৩. দ্বিতীয় ধাপ: সমস্যাটির প্রকারভেদ চিহ্নিত করা (**Classification** থেকে **Regression**-এ পরিবর্তন)

এখানে ML সমস্যার প্রকৃতি চিহ্নিত করা হয় এবং একটি গুরুত্বপূর্ণ পরিবর্তন ঘটে:

- মূল টাস্ক: চর্ন রেট কমানোর জন্য প্রধান কাজ হলো সেই গ্রাহকদের চিহ্নিত করা যারা প্ল্যাটফর্ম ছেড়ে চলে যাবেন। চিহ্নিত হওয়ার পর তাদের ডিসকাউন্ট বা অন্যান্য অফার দিয়ে আটকানো যেতে পারে।
- প্রাথমিক বিবেচনা (**Classification**): প্রাথমিকভাবে মনে হয়েছিল এটি একটি ক্লাসিফিকেশন প্রবলেম [৯]। অর্থাৎ, মডেল প্রতিটি গ্রাহকের জন্য অনুমান করবে: "এই মাসে কি সে প্ল্যাটফর্ম ছাড়বে? হ্যাঁ (Yes) নাকি না (No)?"।
- পার্থক্য সৃষ্টি (**The Pivot**): আলোচনার পর দেখা গেল যে, সব গ্রাহককে সমানভাবে ডিসকাউন্ট দেওয়া ব্যয়বহুল [১০]। বরং, যাদের প্ল্যাটফর্ম ছাড়ার সম্ভাবনা বেশি, তাদের বেশি ডিসকাউন্ট দেওয়া উচিত, আর যাদের কম সম্ভাবনা, তাদের কম ডিসকাউন্ট দেওয়া উচিত।
- চূড়ান্ত রূপ (**Regression**): তাই, মডেল থেকে শুধু 'হ্যাঁ/না' উত্তর না চেয়ে, গ্রাহকের প্ল্যাটফর্ম ছেড়ে যাওয়ার সম্ভাবনা (**Probability**) একটি স্কেল আকারে (০ থেকে ১০০%) জানতে চাওয়া হলো। যেহেতু আউটপুট একটি অবিচ্ছিন্ন মান (Continuous Score), তাই সমস্যাটি ক্লাসিফিকেশন থেকে রিগ্রেশন প্রবলেম-এ রূপান্তরিত হলো।

৪. তৃতীয় ও চতুর্থ ধাপ: বিদ্যমান সমাধান ও ডেটা প্রয়োজনীয়তা

মডেল তৈরি শুরুর আগে ডেটা সায়েন্টিস্টকে দুটি কাজ করতে হয়:

- বিদ্যমান সমাধানের সন্ধান: খোঁজ নেওয়া উচিত যে এই সমস্যার ওপর আগে কেউ কাজ করেছে কি না (যেমন, কোনো পুরনো চর্ন রেট প্রেডিকশন মডেল)। যদি থাকে, তবে সেই মডেল থেকে ব্যবহৃত ফ্যাক্টর বা ফিচার (features) সম্পর্কে ধারণা পাওয়া যেতে পারে, যা প্রজেক্টে সাহায্য করবে।
- ডেটা ফিচার নির্ধারণ: প্ল্যাটফর্ম ছেড়ে যাওয়ার সম্ভাবনা অনুমান করার জন্য কোন কোন ডেটা কলাম দরকার, তা স্পষ্টভাবে নির্ধারণ করা।
 - প্রয়োজনীয় ফিচার: কত সময় ধরে ব্যবহারকারী নেটফ্লিক্স দেখছেন (ওয়াচ টাইম), কত সময় ব্রাউজিং করছেন (যা বোঝায় তারা তাদের পছন্দের কনটেন্ট পাচ্ছেন না), কতগুলি সার্চ করে ফল পাননি, কতগুলি মুভি বা শো শুরু করে মাঝপথে ছেড়ে দিয়েছেন, এবং কতগুলি রেকমেন্ডেশনে ক্লিক করেছেন।
 - ডেটা ইঞ্জিনিয়ারিং-এর প্রয়োজনীয়তা: ডেটা সায়েন্টিস্ট এই সমস্ত ডেটা সরাসরি ডেটাবেস থেকে আনতে পারেন না। এই কারণে তাঁকে একজন ডেটা ইঞ্জিনিয়ারের সাথে কথা বলতে হয়। ডেটা ইঞ্জিনিয়ার মূল ডেটাবেস থেকে এই প্রয়োজনীয় ডেটাগুলি সংগ্রহ করে একটি ডেটা ওয়ারহাউস তৈরি করে দেন, যার ওপর ডেটা সায়েন্টিস্ট কাজ করেন।

৫. পঞ্চম ধাপ: সাফল্যের মাপকাঠি (Metrics) নির্ধারণ

একটি মডেল সফল হয়েছে কি না, তা বিচার করার জন্য শুরুতেই কিছু ম্যাট্রিক্স তৈরি করতে হয়।

- মূল ম্যাট্রিক্স: প্রজেক্টের শেষে দেখা হয় যে, চর্ন রেট কমানোর লক্ষ্য (যেমন ০.২৫% কমানো) পূরণ হয়েছে কি না।
- অন্যান্য ম্যাট্রিক্স: এটিও দেখা উচিত যে মডেল যাদের প্ল্যাটফর্ম ছাড়ার পূর্বাভাস দিয়েছিল, আসলেই কি তারাই প্ল্যাটফর্ম ছেড়েছেন। ম্যাট্রিক্সগুলি ডেটা সায়েন্টিস্টকে সঠিক পথে পরিচালিত করে।

৬. ষষ্ঠ ধাপ: লার্নিং কৌশল (Online vs. Batch) নির্বাচন

মডেলকে কীভাবে প্রশিক্ষণ দেওয়া হবে, তা সিদ্ধান্ত নেওয়া জরুরি।

- ব্যাচ লার্নিং: মডেল একবার প্রশিক্ষণ দেওয়া হয় (অফলাইন), ডিপ্লয় করা হয়, এবং পুনরায় প্রশিক্ষণের প্রয়োজন হলে নতুন ডেটা সহ আবার অফলাইনে প্রশিক্ষণ দেওয়া হয়।
- অনলাইন লার্নিং: মডেল নতুন ডেটা আসার সাথে সাথে ক্রমাগত শিখতে থাকে।
- সিদ্ধান্ত: নেটফ্লিক্সের ব্যবহারকারীদের আচরণ খুব পরিবর্তনশীল (যেমন, ছুটি বা লকডাউনের সময় বেশি দেখা হয়)। তাই অনলাইন লার্নিং তত্ত্বগতভাবে ভালো হলেও, এটি ডেটা সেটআপের কারণে জটিল হতে পারে। তাই বিকল্প হিসেবে প্রতি সপ্তাহে একবার অফলাইনে মডেলটিকে প্রশিক্ষণ দিয়ে আপডেট করার (রিট্রেনিং) সিদ্ধান্ত নেওয়া হয়।

৭. সপ্তম ধাপ: অনুমানগুলি পরীক্ষা করা (Assumptions Check)

কোডিং শুরু করার আগে শেষ ধাপ হলো প্রজেক্টের সমস্ত অনুমানগুলি যাচাই করা।

- ফিচারগুলির উপলব্ধতা: নিশ্চিত করতে হবে যে ডেটা সায়েন্টিস্ট যে ফিচারগুলি (যেমন ওয়াচ টাইম, সার্চ রেজাল্ট না পাওয়ার সংখ্যা) ব্যবহার করতে চান, তা ডেটাবেসে সত্যিই রয়েছে।
- ভৌগোলিক পরিধি: যদি কোম্পানিটি মাল্টিন্যাশনাল হয়, তবে দেখতে হবে যে একটি মডেলই কি বিশ্বের সব অঞ্চলে সমানভাবে কাজ করবে, নাকি বিভিন্ন ভৌগোলিক অঞ্চলের জন্য আলাদা মডেল তৈরি করতে হবে।

৮. উপসংহার ও পরামর্শ

একজন ডেটা সায়েন্টিস্ট হিসেবে কোডিং শুরু করার আগে সমস্যার প্রতিটি দিক নিয়ে গভীরভাবে চিন্তা করা এবং পরিকল্পনা করা জরুরি। বড় কোম্পানিগুলিতে ভুল পথে কাজ শুরু করলে সময় ও অর্থ নষ্ট হয়, যা লিডারশিপের ওপর নেতিবাচক প্রভাব ফেলে। এই ধরনের কৌশলগত চিন্তাভাবনা এবং নেতৃত্ব দেওয়ার ক্ষমতাই একজন ব্যক্তিকে ম্যানেজার বা সিনিয়র পজিশনে নিয়ে যায়।